(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 July 2001 (05.07.2001)

PCT

(10) International Publication Number
WO 01/47944 A2

(51) International Patent Classification[7]: C07H 21/04, 21/02, C12Q 1/68, C07K 14/47, 16/18, G01N 33/53, A61K 48/00, 39/395, 38/00

(21) International Application Number: PCT/US00/35498

(22) International Filing Date:
28 December 2000 (28.12.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/173,419	28 December 1999 (28.12.1999)	US
Not furnished	27 December 2000 (27.12.2000)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	60/173,419 (CIP)
Filed on	28 December 1999 (28.12.1999)
US	Not furnished (CIP)
Filed on	27 December 2000 (27.12.2000)

(71) Applicant *(for all designated States except US)*: **CURAGEN CORPORATION** [US/US]; 555 Long Wharf Drive, 11th Floor, Branford, CT 06511 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SHIMKETS, Richard, A.** [US/US]; 191 Leete Street, West Haven, CT 06516 (US). **LEACH, Martin** [GB/US]; 884 School Street, Webster, MA 01570 (US).

(74) Agent: **ELRIFI, Ivor, R.**; Mintz, Levin, Cohn, Ferris, Glovsky, and Popeo, PC, One Financial Center, Boston, MA 02111 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

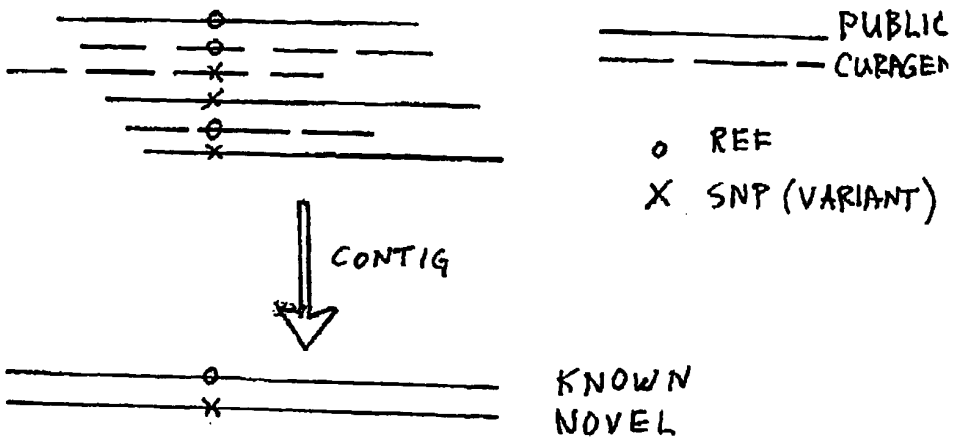
(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: NUCLEIC ACIDS CONTAINING SINGLE NUCLEOTIDE POLYMORPHISMS AND METHODS OF USE THEREOF



(57) Abstract: The invention provides nucleic acids containing single-nucleotide polymorphisms identified for transcribed human sequences, as well as methods of using the nucleic acids.

WO 01/47944 A2

NUCLEIC ACIDS CONTAINING SINGLE NUCLEOTIDE POLYMORPHISMS AND METHODS OF USE THEREOF

BACKGROUND OF THE INVENTION

Sequence polymorphism-based analysis of nucleic acid sequences can augment or replace previously known methods for determining the identity and relatedness of individuals. The approach is generally based on alterations in nucleic acid sequences between related individuals. This analysis has been widely used in a variety of genetic, diagnostic, and forensic applications. For example, polymorphism analyses are used in identity and paternity analysis, and in genetic mapping studies.

One such type of variation is a restriction fragment length polymorphism (RFLP). RFLPS can create or delete a recognition sequence for a restriction endonuclease in one nucleic acid relative to a second nucleic acid. The result of the variation is an alteration in the relative length of restriction enzyme generated DNA fragments in the two nucleic acids.

Other polymorphisms take the form of short tandem repeats (STR) sequences, which are also referred to as variable numbers of tandem repeat (VNTR) sequences. STR sequences typically include tandem repeats of 2, 3, or 4 nucleotide sequences that are present in a nucleic acid from one individual but absent from a second, related individual at the corresponding genomic location.

Other polymorphisms take the form of single nucleotide variations, termed single nucleotide polymorphisms (SNPs), between individuals. A SNP can, in some instances, be referred to as a "cSNP" to denote that the nucleotide sequence containing the SNP originates as a cDNA.

SNPs can arise in several ways. A single nucleotide polymorphism may arise due to a substitution of one nucleotide for another at the polymorphic site. Substitutions can be transitions or transversions. A transition is the replacement of one purine nucleotide by another purine nucleotide, or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine, or the converse.

Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele. Thus, the polymorphic site is a site at which one allele bears a gap with respect to a single nucleotide in another allele. Some SNPs

occur within, or near genes. One such class includes SNPs falling within regions of genes encoding for a polypeptide product. These SNPs may result in an alteration of the amino acid sequence of the polypeptide product and give rise to the expression of a defective or other variant protein. Such variant products can, in some cases result in a pathological condition, *e.g.*, genetic disease. Examples of genes in which a polymorphism within a coding sequence gives rise to genetic disease include sickle cell anemia and cystic fibrosis. Other SNPs do not result in alteration of the polypeptide product. Of course, SNPs can also occur in noncoding regions of genes.

SNPs tend to occur with great frequency and are spaced uniformly throughout the genome. The frequency and uniformity of SNPs means that there is a greater probability that such a polymorphism will be found in close proximity to a genetic locus of interest.

SUMMARY OF THE INVENTION

The invention is based in part on the discovery of novel single nucleotide polymorphisms (SNPs) in regions of human DNA.

Accordingly, in one aspect, the invention provides an isolated polynucleotide which includes one or more of the SNPs described herein. The polynucleotide can be, *e.g.*, a nucleotide sequence which includes one or more of the polymorphic sequences shown in Table 1 and the Sequence Listing (SEQ ID NOS: 1 - 7867) and which includes a polymorphic sequence, or a fragment of the polymorphic sequence, as long as it includes the polymorphic site. The polynucleotide may alternatively contain a nucleotide sequence which includes a sequence complementary to one or more of the sequences (SEQ ID NOS: 1-7867), or a fragment of the complementary nucleotide sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence.

The polynucleotide can be, *e.g.*, DNA or RNA, and can be between about 10 and about 100 nucleotides, *e.g,* 10-90, 10-75, 10-51, 10-40, or 10-30, nucleotides in length.

In some embodiments, the polymorphic site in the polymorphic sequence includes a nucleotide other than the nucleotide listed in Table 1, column 5 for the polymorphic sequence, *e.g.*, the polymorphic site includes the nucleotide listed in Table 1, column 6 for the polymorphic sequence.

In other embodiments, the complement of the polymorphic site includes a nucleotide other than the complement of the nucleotide listed in Table 1, column 5 for the complement of the polymorphic sequence, *e.g.*, the complement of the nucleotide listed in Table 1, column 6 for the polymorphic sequence.

In some embodiments, the polymorphic sequence is associated with a polypeptide related to one of the protein families disclosed herein. For example, the nucleic acid may be associated with a polypeptide related to an ATPase associated protein, a cadherin, or any of the other proteins identified in Table 1, column 10.

In another aspect, the invention provides an isolated allele-specific oligonucleotide that hybridizes to a first polynucleotide containing a polymorphic site. The first polynucleotide can be, *e.g.*, a nucleotide sequence comprising one or more polymorphic sequences (SEQ ID NOS:1 - 7867), provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence. Alternatively, the first polynucleotide can be a nucleotide sequence that is a fragment of the polymorphic sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence, or a complementary nucleotide sequence which includes a sequence complementary to one or more polymorphic sequences (SEQ ID NOS:1 - 7867), provided that the complementary nucleotide sequence includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5. The first polynucleotide may in addition include a nucleotide sequence that is a fragment of the complementary sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence.

In some embodiments, the oligonucleotide does not hybridize under stringent conditions to a second polynucleotide. The second polynucleotide can be, *e.g.*, (a) a nucleotide sequence comprising one or more polymorphic sequences (SEQ ID NOS:1 - 7867), wherein the polymorphic sequence includes the nucleotide listed in Table 1, column 5 for the polymorphic sequence; (b) a nucleotide sequence that is a fragment of any of the polymorphic sequences; (c) a complementary nucleotide sequence including a sequence complementary to one or more polymorphic sequences (SEQ ID NOS:1 - 7867), wherein the polymorphic sequence includes the complement of the nucleotide listed in Table 1, column 5; and (d) a nucleotide sequence that is a fragment of the complementary sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence.

The oligonucleotide can be, *e.g.*, between about 10 and about 100 bases in length. In some embodiments, the oligonucleotide is between about 10 and 75 bases, 10 and 51 bases, 10 and about 40 bases, or about 15 and 30 bases in length.

The invention also provides a method of detecting a polymorphic site in a nucleic acid. The method includes contacting the nucleic acid with an oligonucleotide that hybridizes to a polymorphic sequence selected from the group consisting of SEQ ID NOS: 1-7867, or its complement, provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or the complement includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5. The method also includes determining whether the nucleic acid and the oligonucleotide hybridize. Hybridization of the oligonucleotide to the nucleic acid sequence indicates the presence of the polymorphic site in the nucleic acid.

In preferred embodiments, the oligonucleotide does not hybridize to the polymorphic sequence when the polymorphic sequence includes the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or when the complement of the polymorphic sequence includes the complement of the nucleotide recited in Table 1, column 5 for the polymorphic sequence.

The oligonucleotide can be, *e.g.*, between about 10 and about 100 bases in length. In some embodiments, the oligonucleotide is between about 10 and 75 bases, 10 and 51 bases, 10 and about 40 bases, or about 15 and 30 bases in length.

In some embodiments, the polymorphic sequence identified by the oligonucleotide is associated with a polypeptide related to one of the protein families disclosed herein. For example, the nucleic acid may be associated polypeptide related to an ATPase associated protein, cadherin, or any of the other protein families identified in Table 1, column 10.

In another aspect, the method includes determining if a sequence polymorphism is present in a subject, such as a human. The method includes providing a nucleic acid from the subject and contacting the nucleic acid with an oligonucleotide that hybridizes to a polymorphic sequence selected from the group consisting of SEQ ID NOS: 1-7867, or its complement, provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for said polymorphic sequence, or the complement includes a nucleotide other than the complement of the nucleotide recited in Table 1,

column 5. Hybridization between the nucleic acid and the oligonucleotide is then determined. Hybridization of the oligonucleotide to the nucleic acid sequence indicates the presence of the polymorphism in said subject.

In a further aspect, the invention provides a method of determining the relatedness of a first and second nucleic acid. The method includes providing a first nucleic acid and a second nucleic acid and contacting the first nucleic acid and the second nucleic acid with an oligonucleotide that hybridizes to a polymorphic sequence selected from the group consisting of SEQ ID NOS: 1-7867, or its complement, provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or the complement includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5. The method also includes determining whether the first nucleic acid and the second nucleic acid hybridize to the oligonucleotide, and comparing hybridization of the first and second nucleic acids to the oligonucleotide. Hybridization of first and second nucleic acids to the nucleic acid indicates the first and second subjects are related.

In preferred embodiments, the oligonucleotide does not hybridize to the polymorphic sequence when the polymorphic sequence includes the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or when the complement of the polymorphic sequence includes the complement of the nucleotide recited in Table 1, column 5 for the polymorphic sequence.

The oligonucleotide can be, *e.g.*, between about 10 and about 100 bases in length. In some embodiments, the oligonucleotide is between about 10 and 75 bases, 10 and 51 bases, 10 and about 40 bases, or about 15 and 30 bases in length.

The method can be used in a variety of applications. For example, the first nucleic acid may be isolated from physical evidence gathered at a crime scene, and the second nucleic acid may be obtained from a person suspected of having committed the crime. Matching the two nucleic acids using the method can establish whether the physical evidence originated from the person.

In another example, the first sample may be from a human male suspected of being the father of a child and the second sample may be from the child. Establishing a match using the described method can establish whether the male is the father of the child.

In another aspect, the invention provides an isolated polypeptide comprising a polymorphic site at one or more amino acid residues, and wherein the protein is encoded by a polynucleotide including one of the polymorphic sequences SEQ ID NOS:1-7867, or their complement, provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or the complement includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5.

The polypeptide can be, *e.g.*, related to one of the protein families disclosed herein. For example, the polypeptide can be related to an ATPase associated protein, cadherin, or any of the other proteins provided in Table 1, column 10.

In some embodiments, the polypeptide is translated in the same open reading frame as is a wild type protein whose amino acid sequence is identical to the amino acid sequence of the polymorphic protein except at the site of the polymorphism.

In some embodiments, the polypeptide encoded by the polymorphic sequence, or its complement, includes the nucleotide listed in Table 1, column 6 for the polymorphic sequence, or the complement includes the complement of the nucleotide listed in Table 1, column 6.

The invention also provides an antibody that binds specifically to a polypeptide encoded by a polynucleotide comprising a nucleotide sequence encoded by a polynucleotide selected from the group consisting of polymorphic sequences SEQ ID NOS:1-7867, or its complement. The polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence, or the complement includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5.

In some embodiments, the antibody binds specifically to a polypeptide encoded by a polymorphic sequence which includes the nucleotide listed in Table 1, column 6 for the polymorphic sequence.

Preferably, the antibody does not bind specifically to a polypeptide encoded by a polymorphic sequence which includes the nucleotide listed in Table 1, column 5 for the polymorphic sequence.

The invention further provides a method of detecting the presence of a polypeptide having one or more amino acid residue polymorphisms in a subject. The method includes providing a protein sample from the subject and contacting the sample with the above-described antibody under conditions that allow for the formation of antibody-antigen complexes. The antibody-antigen complexes are then detected. The presence of the complexes indicates the presence of the polypeptide.

The invention also provides a method of treating a subject suffering from, at risk for, or suspected of, suffering from a pathology ascribed to the presence of a sequence polymorphism in a subject, *e.g.*, a human, non-human primate, cat, dog, rat, mouse, cow, pig, goat, or rabbit. The method includes providing a subject suffering from a pathology associated with aberrant expression of a first nucleic acid comprising a polymorphic sequence selected from the group consisting of SEQ ID NOS:1 - 7867, or its complement, and treating the subject by administering to the subject an effective dose of a therapeutic agent. Aberrant expression can include qualitative alterations in expression of a gene, *e.g.*, expression of a gene encoding a polypeptide having an altered amino acid sequence with respect to its wild-type counterpart. Qualitatively different polypeptides can include, shorter, longer, or altered polypeptides relative to the amino acid sequence of the wild-type polypeptide. Aberrant expression can also include quantitative alterations in expression of a gene. Examples of quantitative alterations in gene expression include lower or higher levels of expression of the gene relative to its wild-type counterpart, or alterations in the temporal or tissue-specific expression pattern of a gene. Finally, aberrant expression may also include a combination of qualitative and quantitative alterations in gene expression.

The therapeutic agent can be administered to a subject suffering from a pathology associated with aberrant expression of a first nucleic acid comprising a polymorphic sequence. The therapeutic agent can include, e.g., second nucleic acid comprising the polymorphic sequence, provided that the second nucleic acid comprises the nucleotide present in the wild type allele. In some embodiments, the second nucleic acid sequence comprises a polymorphic sequence which includes the nucleotide listed in Table 1, column 5 for the polymorphic sequence.

Alternatively, the therapeutic agent can be a polypeptide encoded by a polynucleotide comprising a polymorphic sequence selected from the group consisting of SEQ ID NOS:1 - 7867, or by a polynucleotide comprising a nucleotide sequence that is complementary to any

one of the polymorphic sequences SEQ ID NOS:1 - 7867, provided that the polymorphic sequence includes the nucleotide listed in Table 1, column 6 for the polymorphic sequence.

The therapeutic agent may further include an antibody as herein described, or an oligonucleotide comprising a polymorphic sequence selected from the group consisting of SEQ ID NOS:1 - 7867, or by a polynucleotide comprising a nucleotide sequence that is complementary to any one of polymorphic sequences SEQ ID NOS:1 - 7867, provided that the polymorphic sequence includes the nucleotide listed in Table 1, column 5 or Table 1, column 6 for the polymorphic sequence.

In another aspect, the invention provides an oligonucleotide array comprising one or more oligonucleotides hybridizing to a first polynucleotide at a polymorphic site encompassed therein. The first polynucleotide can be, e.g., a nucleotide sequence comprising one or more polymorphic sequences (SEQ ID NOS:1 - 7867); a nucleotide sequence that is a fragment of any of the nucleotide sequences, provided that the fragment includes a polymorphic site in the polymorphic sequence; a complementary nucleotide sequence comprising a sequence complementary to one or more polymorphic sequences (SEQ ID NOS:1 - 7867); or a nucleotide sequence that is a fragment of the complementary sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence.

In preferred embodiments, the array comprises 10; 100; 1,000; 10,000; 100,000 or more oligonucleotides.

The invention also provides a kit comprising one or more of the herein-described nucleic acids. The kit can include, e.g., a polynucleotide which includes one or more of the SNPs described herein. The polynucleotide can be, e.g., a nucleotide sequence which includes one or more of the polymorphic sequences shown in Table 1 and the Sequence Listing (SEQ ID NOS: 1 - 7867) and which includes a polymorphic sequence, or a fragment of the polymorphic sequence, as long as it includes the polymorphic site. The polynucleotide may alternatively contain a nucleotide sequence which includes a sequence complementary to one or more of the sequences (SEQ ID NOS:1-7867), or a fragment of the complementary nucleotide sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence. The invention provides an isolated allele-specific oligonucleotide that hybridizes to a first polynucleotide containing a polymorphic site. The first polynucleotide can be, e.g., a nucleotide sequence comprising one or more polymorphic

sequences (SEQ ID NOS:1 - 7867), provided that the polymorphic sequence includes a nucleotide other than the nucleotide recited in Table 1, column 5 for the polymorphic sequence. Alternatively, the first polynucleotide can be a nucleotide sequence that is a fragment of the polymorphic sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence, or a complementary nucleotide sequence which includes a sequence complementary to one or more polymorphic sequences (SEQ ID NOS:1 - 7867), provided that the complementary nucleotide sequence includes a nucleotide other than the complement of the nucleotide recited in Table 1, column 5. The first polynucleotide may in addition include a nucleotide sequence that is a fragment of the complementary sequence, provided that the fragment includes a polymorphic site in the polymorphic sequence.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In the case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description and claims.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides human SNPs in sequences which are transcribed, *i.e.*, are cSNPs. Many SNPs have been identified in genes related to polypeptides of known function. If desired, SNPs associated with various polypeptides can be used together. For example, SNPs can be grouped according to whether they are derived from a nucleic acid encoding a polypeptide related to particular protein family or involved in a particular function. Similarly, SNPs can be grouped according to the functions played by their gene products. Such functions include, structural proteins, proteins which are associated with metabolic pathways, including fatty acid metabolism, glycolysis, intermediary metabolism, calcium metabolism, proteases, and amino acid metabolism, etc. Specifically, the present invention

provides a large number of human cSNP's based on at least one gene product that has not been previously identified. In contrast, and as defined specifically in the following paragraph, the cSNP's involve nucleic acid sequences that are assembled from at least one known sequence.

The present invention provides a large number of human cSNP's based on at least one gene product that has not been previously identified. In contrast, and as defined specifically in the following paragraph, the cSNP's involve nucleic acid sequences that are assembled from at least one known sequence.

7867 distinct polymorphic sites were identified by the present inventors, using the following procedure. Raw traces underlying sequence data were drawn from public databases and from the proprietary database of the Assignee of the present invention. The sequences were obtained by calling the bases from these traces, and included assigning "Phred" quality scores for each called base. For each allelic set, at the polynucleotide level, four or more nucleotide sequences were identified having at least partial overlap with one another.

As illustrated in FIG. 1, these four or more sequences could be clustered and assembled to make a consensus contig that included an ORF. In this way, the inventors found that the assembled contigs defined associated sets of two, or possibly more than two, alleles defined by an SNP at a particular polymorphic site. In order to be confirmed as a SNP site, the nucleotide change from the consensus sequence had to occur in at least two individual sequences, and had to have a "Phred" score of 23 or higher at the site of the presumed SNP. Furthermore, in a window of 5 bases on either side of the SNP, no more than 50% mismatching with the consensus sequence was allowed. In the assembly leading to each of the contigs defining the allelic set, the SNP alleles occur in polynucleotides found in public databases. Furthermore, it was found that the assembled contigs defined associated sets of two, or possibly more than two, alleles defined by an SNP at a particular polymorphic site. These associations were not previously known. The SNPs are presented in Table 1.

At the level of translation of an ORF contained in the contigs, however, the inventors identified allelic sets in which one allele defines a known polypeptide sequence that includes the polymorphic site and another polypeptide allele is not previously known. Then, various associations of alleles are possible. For example, it is possible that an allelic pair is defined

in a noncoding region of the contig containing an ORF. In such cases the inventors believe that the invention resides in the recognition of the allelic pair; this association has not heretofore been made. Alternatively, sets of allelic contigs may exist in which the polymorphic site is within an ORF, but does not result in an amino acid change among the allelic polypeptides. Here too it is believed that the invention resides in the recognition of the allelic pair; and that this association has not heretofore been made. In yet another alternative, the polymorphic site resides within an ORF and results in an amino acid change, or a frameshift, among the alleles of the allelic set. In the sets of gene products that fall within this group, at least one of the alleles at the polypeptide level is a known protein. At least one of the remaining allele or alleles in the set, carrying a variant amino acid at the polymorphic site, is a novel polypeptide not heretofore known. The invention resides at least in the recognition of the polymorphic allele as being a variant of the known reference polypeptide.

Table 1 provides information concerning the allelic sequences. One of the sequences may be termed a reference polymorphic sequence, and the corresponding second sequence includes the variant SNP at the polymorphic site. Since the reference polypeptide sequence is already known, the Sequence Listing accompanying this application provides only the sequence of the polymorphic allele, while its SEQ ID NO is provided in the Table. A reference to the SEQ ID NO that corresponds to the translated amino acid sequence is also given. The Table includes thirteen columns that provide descriptive information for each cSNP, each of which occupies one row in the Table. The column headings, and a description of each, are given below.

SNPs disclosed in Table 1 were detected by aligning large numbers of sequences from genetically diverse sources of publicly available mRNA libraries (Clontech). Software designed specifically to look for multiple examples of variant bases differing from a consensus sequence was created and deployed. A criteria of a minimum of 2 occurrences of a sequence differing from the consensus in high quality sequence reads was used to identify an SNP.

The SNPs described herein may be useful in diagnostic kits, for DNA arrays on chips and for other uses that involve hybridization of the SNP.

Specific SNPs may have utility where a disease has already been associated with that gene. Examples of possible disease correlations between the claimed SNPs with members of the genes of each classification are listed below:

Amylases

Amylase is responsible for endohydrolysis of 1,4-alpha-glucosidic linkages in oligosaccharides and polysaccharides. Variations in amylase gene may be indicative of delayed maturation and of various amylase producing neoplasms and carcinomas.

Amyloid

The serum amyloid A (SAA) proteins comprise a family of vertebrate proteins that associate predominantly with high density lipoproteins (HDL). The synthesis of certain members of the family is greatly increased in inflammation. Prolonged elevation of plasma SAA levels, as in chronic inflammation, 15 results in a pathological condition, called amyloidosis, which affectsthe liver, kidney and spleen and which is characterized by the highly insoluble accumulation of SAA in these tissues. Amyloid selectively inhibits insulin-stimulated glucose utilization and glycogen deposition in muscle, while not affecting adipocyte glucose metabolism. Deposition of fibrillar amyloid proteins intraneuronally, as neurofibrillary tangles, extracellularly, as plaques and in blood vessels, is characteristic of both Alzheimer's disease and aged Down's syndrome. Amyloid deposition is also associated with type II diabetes mellitus.

Angiopoeitin

Members of the angiopoeitin/fibrinogen family have been shown to stimulate the generation of new blood vessels, inhibit the generation of new blood vessels, and perform several roles in blood clotting. This generation of new blood vessels, called angiogenesis, is also an essential step in tumor growth in order for the tumor to get the blood supply it needs to expand. Variation in these genes may be predictive of any form of heart disease, numerous blood clotting disorders, stroke, hypertension and predisposition to tumor formation and metastasis. In particular, these variants may be predictive of the response to various antihypertensive drugs and chemotherapeutic and anti-tumor agents.

Apoptosis-related proteins

Active cell suicide (apoptosis) is induced by events such as growth factor withdrawal and toxins. It is controlled by regulators, which have either an inhibitory effect on programmed cell death (anti-apoptotic) or block the protective effect of inhibitors (pro-apoptotic). Many viruses have found a way of countering defensive apoptosis by encoding their own anti-apoptosis genes preventing their target-cells from dying too soon. Variants of apoptosis related genes may be useful in formulation of antiaging drugs.

Cadherin, Cyclin, Polymerase, Oncogenes, Histones, Kinases

Members of the cell division/cell cycle pathways such as cyclins, many transcription factors and kinases, DNA polymerases, histones, helicases and other oncogenes play a critical role in carcinogenesis where the uncontrolled proliferation of cells leads to tumor formation and eventually metastasis. Variation in these genes may be predictive of predisposition to any form of cancer, from increased risk of tumor formation to increased rate of metastasis. In particular, these variants may be predictive of the response to various chemotherapeutic and anti-tumor agents.

Colony-stimulating factor-related proteins

Granulocyte/macrophage colony-stimulating factors are cytokines that act in hematopoiesis by controlling the production, differentiation, and function of 2 related white cell populations of the blood, the granulocytes and the monocytes-macrophages.

Complement-related proteins

Complement proteins are immune associated cytotoxic agents, acting in a chain reaction to exterminate target cells to that were opsonized (primed) with antibodies, by forming a membrane attack complex (MAC). The mechanism of killing is by opening pores in the target cell membrane. Variations in 20 complement genes or their inhibitors are associated with many autoimmune disorders. Modified serum levels of complement products cause edemas of various tissues, lupus (SLE), vasculitis, glomerulonephritis, renal failure, hemolytic anemia, thrombocytopenia, and arthritis. They interfere with mechanisms of ADCC (antibody dependent cell cytotoxicity), severely impair immune competence and reduce phagocytic ability. Variants of complement genes may also be indicative of type I

diabetes mellitus, meningitis neurological disorders such as Nemaline myopathy, Neonatal hypotonia, muscular disorders such as congenital myopathy and other diseases.

Cytochrome

The respiratory chain is a key biochemical pathway which is essential to all aerobic cells. There are five different cytochromes involved in the chain. These are heme bound proteins which serve as electron carriers. Modifications in these genes may be predictive of ataxia areflexia, dementia and myopathic and neuropathic changes in muscles. Also, association with various types of solid tumors.

Kinesins

Kinesins are tubulin molecular motors that function to transport organelles within cells and to move chromosomes along microtubules during cell division. Modifications of these genes may be indicative of neurological disorders such as Pick disease of the brain, tuberous sclerosis.

Cytokines, Interferon, Interleukin

Members of the cytokine families are known for their potent ability to stimulate cell growth and division even at low concentrations. Cytokines such as erythropoietin are cell-specific in their growth stimulation; erythropoietin is useful for the stimulation of the proliferation of erythroblasts. Variants in cytokines may be predictive for a wide variety of diseases, including cancer predisposition.

G-protein coupled receptors

G-protein coupled receptors (also called R7G) are an extensive group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. Alterations in genes coding for G-coupled proteins may be involved in and indicative of a vast number of physiological conditions. These include blood pressure regulation, renal dysfunctions, male infertility, dopamine associated cognitive, emotional, and endocrine functions, hypercalcemia, chondrodysplasia and osteoporosis, pseudohypoparathyroidism, growth retardation and dwarfism.

Thioesterases

Eukaryotic thiol proteases are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. Variants of thioester associated genes may be predictive of neuronal disorders and mental illnesses such as Ceroid Lipoffiscinosis, Neuronal 1, Infantile, Santavuori disease and more.

The SNPs are shown in Table 1 and the Sequence Listing. Both provide a summary of the polymorphic sequences disclosed herein. In the Table, a "SNP" is a polymorphic site embedded in a polymorphic sequence. The polymorphic site is occupied by a single nucleotide, which is the position of nucleotide variation between the wild type and polymorphic allelic sequences. The site is usually preceded by and followed by relatively highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations). Thus, a polymorphic sequence can include one or more of the following sequences: (1) a sequence having the nucleotide denoted in Table 1, column 5 at the polymorphic site in the polymorphic sequence; or (2) a sequence having a nucleotide other than the nucleotide denoted in Table 1, column 5 at the polymorphic site in the polymorphic sequence. An example of the latter sequence is a polymorphic sequence having the nucleotide denoted in Table 1, column 6 at the polymorphic site in the polymorphic sequence.

Nucleotide sequences for a referenced-polymorphic pair are presented in Table 1. Each cSNP entry provides information concerning the wild type nucleotide sequence as well as the corresponding sequence that includes the SNP at the polymorphic site. Since the wild type sequence is already known, the Sequence Listing accompanying this application provides only the sequence of the polymorphic allele; its SEQ ID NO: is also cross referenced in the Table 1. A reference to the SEQ ID NO: giving the translated amino acid sequence is also given if appropriate. The Table includes thirteen columns that provide descriptive information for each cSNP, each of which occupies one row in the Table. The column headings, and an explanation for each, are given below.

"SEQ ID" provides the cross-references to the nucleotide SEQ ID NOs: for the polymorphic sequences, which are numbered consecutively, and, as explained below, amino acid SEQ ID NOs: as well, in the Sequence Listing of the application. Conversely, each

sequence entry in the Sequence Listing also includes a cross-reference to the CuraGen sequence ID, under the label "CuraGen sequence ID". The first SEQ ID NO: given in the first column of each row of the Table is the SEQ ID NO: identifying the nucleic acid sequence for the polymorphisms. If a polymorphism carries an entry for an amino acid in a coding region, then a second SEQ ID NO: appears in parentheses in the column "Amino acid after" (see below) for the polymorphic amino acid sequence . The latter SEQ ID NOs: refer to amino acid sequences giving the polymorphic amino acid sequences that are the translation of the nucleotide polymorphism. If a polymorphism carries no entry for the protein portion of the row, only one SEQ ID NO: is provided, in the first column.

"Base pos. of SNP" gives the numerical position of the nucleotide in the nucleic acid at which the cSNP is found, as identified in this invention.

"Polymorphic sequence" provides a 51-base sequence with the polymorphic site at the 26th base in the sequence, as well as 25 bases from the reference sequence on the 5' side and the 3' side of the polymorphic site. The designation at the polymorphic site is enclosed in square brackets, and provides first, the reference nucleotide; second, a "slash (/)"; and third, the polymorphic nucleotide. In certain cases the polymorphism is an insertion or a deletion. In that case, the position which is "unfilled" (i.e., the reference or the polymorphic position) is indicated by the word "gap".

"Base before" provides the nucleotide present in the reference sequence at the position at which the polymorphism is found.

"Base after" provides the altered nucleotide at the position of the polymorphism.

"Amino acid before" provides the amino acid in the reference protein, if the polymorphism occurs in a coding region.

"Amino acid after" provides the amino acid in the polymorphic protein, if the polymorphism occurs in a coding region. This column also includes the SEQ ID NO: in parentheses for the translated polymorphic amino acid sequence if the polymorphism occurs in a coding region.

"Type of change" provides information on the nature of the polymorphism.

"SILENT-NONCODING" is used if the polymorphism occurs in a noncoding region of a nucleic acid.

"SILENT-CODING" is used if the polymorphism occurs in a coding region of a nucleic acid of a nucleic acid and results in no change of amino acid in the translated polymorphic protein.

"CONSERVATIVE" is used if the polymorphism occurs in a coding region of a nucleic acid and provides a change in which the altered amino acid falls in the same class as the reference amino acid. The classes are:

Aliphatic: Gly, Ala, Val, Leu, Ile;

Aromatic: Phe, Tyr, Trp;

Sulfur-containing: Cys, Met;

Aliphatic OH: Ser, Thr;

Basic: Lys, Arg, His;

Acidic: Asp, Glu, Asn, Gln;

Pro falls in none of the other classes; and

End defines a termination codon.

"NONCONSERVATIVE" is used if the polymorphism occurs in a coding region of a nucleic acid and provides a change in which the altered amino acid falls in a different class than the reference amino acid.

"FRAMESHIFT" relates to an insertion or a deletion. If the frameshift occurs in a coding region, the Table provides the translation of the frameshifted codons 3' to the polymorphic site.

"Protein classification of CuraGen gene" provides a generic class into which the protein is classified. During the course of the work leading to the filing of the four applications identified above, approximately 100 classes of proteins were identified.

"Name of protein identified following a BLASTX analysis of the CuraGen sequence" provides the database reference for the protein found to resemble the novel reference-polymorphism cognate pair most closely. (The next paragraph explains how a sequence was determined to be "novel").

"Similarity (pvalue) following a BLASTX analysis" provides the pvalue, a statistical measure from the BLASTX analysis that the polymorphic sequence is similar to, and therefore an allele of, the reference, or wild-type, sequence. In the present application, a cutoff of pvalue > 1×10^{-50} (entered, for example, as 1.0E-50 in the Table) is used to establish that the reference-polymorphic cognate pairs are novel.

"Map location" provides any information available at the time of filing related to localization of a gene on a chromosome.

The polymorphisms are arranged in the Table in the following order.

SEQ ID NOs: 1-5696 are nucleotide sequences for SNPs that are silent.

SEQ ID NOs: 5697-6011 are nucleotide sequences for SNPs that lead to conservative amino acid changes.

SEQ ID NOs: 6012-6740 are nucleotide sequences for SNPs that lead to nonconservative amino acid changes.

SEQ ID NOs: 6741-7867 are nucleotide sequences for SNPs that involve a gap. With respect to the reference or wild-type sequence at the position of the polymorphism, the allelic cSNP introduces an additional nucleotide (an insertion) or deletes a nucleotide (a deletion). An SNP that involves a gap generates a frame shift.

SEQ ID NOs: 7868-8182 are the amino acid sequences centered at the polymorphic amino acid residue for the protein products provided by SNPs that lead to conservative amino acid changes. 7 or 8 amino acids on either side of the polymorphic site are shown. The order in which these sequences appear mirrors the order of presentation of the cognate nucleotide sequences, and is set forth in the Table.

SEQ ID NOs: 8183-8911 are the amino acid sequences centered at the polymorphic amino acid residue for the protein products provided by SNPs that lead to nonconservative

amino acid changes. 7 or 8 amino acids on either side of the polymorphic site are shown. The order in which these sequences appear mirrors the order of presentation of the cognate nucleotide sequences, and is set forth in the Table.

SEQ ID NOs: 8912-10038 are the amino acid sequences centered at the polymorphic amino acid residue for the protein products provided by SNPs that lead to frameshift-induced amino acid changes. 7 or 8 amino acids on either side of the polymorphic site are shown. The order in which these sequences appear mirrors the order of presentation of the cognate nucleotide sequences, and is set forth in the Table.

Provided herein are compositions which include, or are capable of detecting, nucleic acid sequences having these polymorphisms, as well as methods of using nucleic acids.

IDENTIFICATION OF INDIVIDUALS CARRYING SNPS

Individuals carrying polymorphic alleles of the invention may be detected at either the DNA, the RNA, or the protein level using a variety of techniques that are well known in the art. Strategies for identification and detection are described in *e.g.*, EP 730,663, EP 717,113, and PCT US97/02102. The present methods usually employ pre-characterized polymorphisms. That is, the genotyping location and nature of polymorphic forms present at a site have already been determined. The availability of this information allows sets of probes to be designed for specific identification of the known polymorphic forms.

Many of the methods described below require amplification of DNA from target samples. This can be accomplished by e.g., PCR. See generally PCR Technology: Principles and Applications for DNA Amplification (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); PCR Protocols: A Guide to Methods and Applications (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., Nucleic Acids Res. 19, 4967 (1991); Eckert et al., PCR Methods and Applications 1, 17 (1991); PCR (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202.

The phrase "recombinant protein" or "recombinantly produced protein" refers to a peptide or protein produced using non-native cells that do not have an endogenous copy of DNA able to express the protein. In particular, as used herein, a recombinantly produced protein relates to the gene product of a polymorphic allele, i.e., a "polymorphic protein"

containing an altered amino acid at the site of translation of the nucleotide polymorphism. The cells produce the protein because they have been genetically altered by the introduction of the appropriate nucleic acid sequence. The recombinant protein will not be found in association with proteins and other subcellular components normally associated with the cells producing the protein. The terms "protein" and "polypeptide" are used interchangeably herein.

The phrase "substantially purified" or "isolated" when referring to a nucleic acid, peptide or protein, means that the chemical composition is in a milieu containing fewer, or preferably, essentially none, of other cellular components with which it is naturally associated. Thus, the phrase "isolated" or "substantially pure" refers to nucleic acid preparations that lack at least one protein or nucleic acid normally associated with the nucleic acid in a host cell. It is preferably in a homogeneous state although it can be in either a dry or aqueous solution. Purity and homogeneity are typically determined using analytical chemistry techniques such as gel electrophoresis or high performance liquid chromatography. Generally, a substantially purified or isolated nucleic acid or protein will comprise more than 80% of all macromolecular species present in the preparation. Preferably, the nucleic acid or protein is purified to represent greater than 90% of all macromolecular species present. More preferably the nucleic acid or protein is purified to greater than 95%, and most preferably the nucleic acid or protein is purified to essential homogeneity, wherein other macromolecular species are not detected by conventional analytical procedures.

The genomic DNA used for the diagnosis may be obtained from any nucleated cells of the body, such as those present in peripheral blood, urine, saliva, buccal samples, surgical specimen, and autopsy specimens. The DNA may be used directly or may be amplified enzymatically in vitro through use of PCR (Saiki et al. Science 239:487-491 (1988)) or other in vitro amplification methods such as the ligase chain reaction (LCR) (Wu and Wallace Genomics 4:560-569 (1989)), strand displacement amplification (SDA) (Walker et al. Proc. Natl. Acad. Sci. U.S.A, 89:392-396 (1992)), self-sustained sequence replication (3SR) (Fahy et al. PCR Methods P&J& 1:25-33 (1992)), prior to mutation analysis.

The method for preparing nucleic acids in a form that is suitable for mutation detection is well known in the art. A "nucleic acid" is a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, including known analogs of natural nucleotides unless otherwise indicated. The term "nucleic acids", as used herein,

refers to either DNA or RNA. "Nucleic acid sequence" or "polynucleotide sequence" refers to a single-stranded sequence of deoxyribonucleotide or ribonucleotide bases read from the 5' end to the 3' end. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are beyond the 5' end of the RNA transcript in the 5' direction are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are beyond the 3' end of the RNA transcript in the 3' direction are referred to as "downstream sequences". The term includes both self-replicating plasmids, infectious polymers of DNA or RNA and nonfunctional DNA or RNA. The complement of any nucleic acid sequence of the invention is understood to be included in the definition of that sequence. "Nucleic acid probes" may be DNA or RNA fragments.

The detection of polymorphisms in specific DNA sequences, can be accomplished by a variety of methods including, but not limited to, restriction-fragment-length-polymorphism detection based on allele-specific restriction-endonuclease cleavage (Kan and Dozy Lancet ii:910-912 (1978)), hybridization with allele-specific oligonucleotide probes (Wallace et al. Nucl. Acids Res. 6:3543-3557 (1978)), including immobilized oligonucleotides (Saiki et al. Proc. Natl. Acad. SCI. USA, 86:6230-6234 (1969)) or oligonucleotide arrays (Maskos and Southern Nucl. Acids Res 21:2269-2270 (1993)), allele-specific PCR (Newton et al. Nucl Acids Res 17:2503-_2516 (1989)), mismatch-repair detection (MRD) (Faham and Cox . Genome Res 5:474-482 (1995)), binding of MutS protein (Wagner et al. Nucl Acids Res 23:3944-3948 (1995), denaturing-gradient gel electrophoresis (DGGE) (Fisher and Lerman et al. *Proc. NatI. Acad. Sci. U.S.A. 80:1579-l* 583 (1983)), single-strand-conformation-polymorphism detection (Orita et al. Genomics 5:874-879 (1983)), RNAase cleavage at mismatched base-pairs (Myers et al. Science 230:1242 (1985)), chemical (Cotton et al. Proc. Natl. w Sci. U.S.A, 8Z4397-4401 (1988)) or enzymatic (Youil et al. Proc. Natl. Acad. Sci. U.S.A. 92:87-91 (1995)) cleavage of heteroduplex DNA, methods based on allele specific primer_extension (Syvanen et al. Genomics 8:684-692 (1990)), genetic bit analysis (GBA) (Nikiforov et al. &&I Acids 22:4167-4175 (1994)), the oligonucleotide-ligation assay (OLA) (Landegren et al. Science_241:1077 (1988)), the allele-specific ligation chain reaction (LCR) (Barrany Proc. Natl. Acad. Sci. U.S.A. 88:189-1 93 (1991)), gap-LCR (Abravaya et al. Nucl Acids Res 23:675-682 (1995)), radioactive and/or fluorescent DNA sequencing using standard procedures well known in the art, and peptide nucleic acid (PNA) assays (Orum et al., Nucl. Acids Res, 21:5332-5356 (1993); Thiede et al., Nucl. Acids Res. 24:983-984

(1996)).

"Specific hybridization" or "selective hybridization" refers to the binding, or duplexing, of a nucleic acid molecule only to a second particular nucleotide sequence to which the nucleic acid is complementary, under suitably stringent conditions when that sequence is present in a complex mixture (e.g., total cellular DNA or RNA). "Stringent conditions" are conditions under which a probe will hybridize to its target subsequence, but to no other sequences. Stringent conditions are sequence-dependent and are different in different circumstances. Longer sequences hybridize specifically at higher temperatures than shorter ones. Generally, stringent conditions are selected such that the temperature is about 5°C lower than the thermal melting point (Tm) for the specific sequence to which hybridization is intended to occur at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength, pH, and nucleic acid concentration) at which 50% of the target sequence hybridizes to the complementary probe at equilibrium. Typically, stringent conditions include a salt concentration of at least about 0.01 to about 1.0 M Na ion concentration (or other salts), at pH 7.0 to 8.3. The temperature is at least about 30°C for short probes (e.g., 10 to 50 nucleotides) . Stringent conditions can also be achieved with the addition of destabilizing agents such as formamide. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridization.

"Complementary" or "target" nucleic acid sequences refer to those nucleic acid sequences which selectively hybridize to a nucleic acid probe. Proper annealing conditions depend, for example, upon a probe's length, base composition, and the number of mismatches and their position on the probe, and must often be determined empirically. For discussions of nucleic acid probe design and annealing conditions, see, for example, Sambrook et al., or Current Protocols in Molecular Biology, F. Ausubel *et* al., ed., Greene Publishing and Wiley-Interscience, New York (1987).

A perfectly matched probe has a sequence perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion of the target sequence. A "polymorphic" marker or site is the locus at which a sequence difference occurs with respect to a reference sequence. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple

sequence repeats, and insertion elements such as Alu. The reference allelic form may be, for example, the most abundant form in a population, or the first allelic form to be identified, and other allelic forms are designated as alternative, variant or polymorphic alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the "wild type" form, and herein may also be referred to as the "reference" form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two distinguishable forms (i.e., base sequences), and a triallelic polymorphism has three such forms.

As used herein an "oligonucleotide" is a single-stranded nucleic acid ranging in length from 2 to about 60 bases. Oligonucleotides are often synthetic but can also be produced from naturally occurring polynucleotides. A probe is an oligonucleotide capable of binding to a target nucleic acid of a complementary sequence through one or more types of chemical bonds, usually through complementary base pairing via hydrogen bond formation. Oligonucleotides probes are often between 5 and 60 bases, and, in specific embodiments, may be between 10-40, or 15-30 bases long. An oligonucleotide probe may include natural (i.e. A, G, C, or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in an oligonucleotide probe may be joined by a linkage other than a phosphodiester bond, such as a phosphoramidite linkage or a phosphorothioate linkage, or they may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than by phosphodiester bonds, so long as it does not interfere with hybridization.

As used herein, the term "primer" refers to a single-stranded oligonucleotide which acts as a point of initiation of template-directed DNA synthesis under appropriate conditions (e.g., in the presence of four different nucleoside triphosphates and a polymerization agent, such as DNA polymerase, RNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. The appropriate length of a primer depends on the intended use of the primer, but typically ranges from 15 to 30 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not be perfectly complementary to the exact sequence of the template, but should be sufficiently complementary to hybridize with it. The term "primer site" refers to the sequence of the target DNA to which a primer hybridizes. The term "primer pair" refers to a set of primers including a 5' (upstream) primer that hybridizes with the 5' end of the DNA sequence to be amplified and a 3' (downstream) primer that hybridizes with the complement of the 3' end of the sequence to be amplified.

DNA fragments can be prepared, for example, by digesting plasmid DNA, or by use of PCR. Oligonucleotides for use as primers or probes are chemically synthesized by methods known in the field of the chemical synthesis of polynucleotides, including by way of non-limiting example the phosphoramidite method described by Beaucage and Carruthers, Tetrahedron Lett 22:1859-1 862 (1981) and the triester method provided by Matteucci, et al., J. Am. Chem. Soc., 103:3185 (1981) both incorporated herein by reference. These syntheses may employ an automated synthesizer, as described in Needham-VanDevanter, D.R., et al., Nucleic Acids Res. 12:61596168 (1984). Purification of oligonucleotides may be carried out by either native acrylamide gel electrophoresis or by anion-exchange HPLC as described in Pearson, J.D. and Regnier, F.E., ,J. Chrom,, 255:137-149 (1983). A double stranded fragment may then be obtained, if desired, by annealing appropriate complementary single strands together under suitable conditions or by synthesizing the complementary strand using a DNA polymerase with an appropriate primer sequence. Where a specific sequence for a nucleic acid probe is given, it is understood that the complementary strand is also identified and included. The complementary strand will work equally well in situations where the target is a double-stranded nucleic acid.

The sequence of the synthetic oligonucleotide or of any nucleic acid fragment can be can be obtained using either the dideoxy chain termination method or the Maxam-Gilbert method (see Sambrook et al. Molecular Cloning - a Laboratory Manual (2nd Ed.), Vols. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, (1989), which is incorporated herein by reference. This manual is hereinafter referred to as "Sambrook et al." ; Zyskind et al., (1988)). Recombinant DNA Laboratory Manual, (Acad. Press, New York). Oligonucleotides useful in diagnostic assays are typically at least 8 consecutive nucleotides in length, and may range upwards of 18 nucleotides in length to greater than 100 or more consecutive nucleotides.

Another aspect of the invention pertains to isolated antisense nucleic acid molecules that are hybridizable to or complementary to the nucleic acid molecule comprising the SNP-containing nucleotide sequences of the invention, or fragments, analogs or derivatives thereof. An "antisense" nucleic acid comprises a nucleotide sequence that is complementary to a "sense" nucleic acid encoding a protein, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. In specific

aspects, antisense nucleic acid molecules are provided that comprise a sequence complementary to at least about 10, about 25, about 50, or about 60 nucleotides or an entire SNP coding strand, or to only a portion thereof.

In one embodiment, an antisense nucleic acid molecule is antisense to a "coding region" of the coding strand of a polymorphic nucleotide sequence of the invention. The term "coding region" refers to the region of the nucleotide sequence comprising codons which are translated into amino acid. In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence of the invention. The term "noncoding region" refers to 5' and 3' sequences which flank the coding region that are not translated into amino acids (*i.e.*, also referred to as 5' and 3' untranslated regions).

Given the coding strand sequences disclosed herein, antisense nucleic acids of the invention can be designed according to the rules of Watson and Crick or Hoogsteen base pairing. For example, the antisense nucleic acid molecule can generally be complementary to the entire coding region of an mRNA, but more preferably as embodied herein, it is an oligonucleotide that is antisense to only a portion of the coding or noncoding region of the mRNA. An antisense oligonucleotide can range in length between about 5 and about 60 nucleotides, preferably between about 10 and about 45 nucleotides, more preferably between about 15 and 40 nucleotides, and still more preferably between about 15 and 30 in length. An antisense nucleic acid of the invention can be constructed using chemical synthesis or enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used.

Examples of modified nucleotides that can be used to generate the antisense nucleic acid include: 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine,

7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid (v), 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, (acp3)w, and 2,6-diaminopurine. Alternatively, the antisense nucleic acid can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (*i.e.*, RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following section).

The antisense nucleic acid molecules of the invention are typically administered to a subject or generated *in situ* such that they hybridize with or bind to cellular mRNA and/or genomic DNA encoding a polymorphic protein to thereby inhibit expression of the protein, *e.g.*, by inhibiting transcription and/or translation. The hybridization can be by conventional nucleotide complementary to form a stable duplex, or, for example, in the case of an antisense nucleic acid molecule that binds to DNA duplexes, through specific interactions in the major groove of the double helix. An example of a route of administration of antisense nucleic acid molecules of the invention includes direct injection at a tissue site. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then administered systemically. For example, for systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, *e.g.*, by linking the antisense nucleic acid molecules to peptides or antibodies that bind to cell surface receptors or antigens. The antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. To achieve sufficient intracellular concentrations of antisense molecules, vector constructs in which the antisense nucleic acid molecule is placed under the control of a strong pol II or pol III promoter are preferred.

In yet another embodiment, the antisense nucleic acid molecule of the invention is an α-anomeric nucleic acid molecule. An α-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual -units, the strands run parallel to each other (Gaultier *et al.* (1987) *Nucleic Acids Res* 15: 6625-6641). The antisense nucleic acid molecule can also comprise a 2'-o-methylribonucleotide (Inoue *et*

al. (1987) *Nucleic Acids Res* 15: 6131-6148) or a chimeric RNA -DNA analogue (Inoue *et al.* (1987) *FEBS Lett* 215: 327-330).

The following terms are used to describe the sequence relationships between two or more nucleic acids or polynucleotides: "reference sequence", "comparison window", "sequence identity", "percentage of sequence identity", and "substantial identity". A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence. Optimal alignment of sequences for aligning a comparison window may, for example, be conducted by the local homology algorithm of Smith and Waterman Adv. Appl. Math, 2482 (1981), by the homology alignment algorithm of Needleman and Wunsch J. Mol. Biol. 48:443 (1970), by the search for similarity method of Pearson and Lipman Proc. Natl. Acad. Sci. U.S.A. 852444 (1988), or by computerized implementations of these algorithms (for example, GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI).

Techniques for nucleic acid manipulation of the nucleic acid sequences harboring the cSNP's of the invention, such as subcloning nucleic acid sequences encoding polypeptides into expression vectors, labeling probes, DNA hybridization, and the like, are described generally in Sambrook et al., The phrase "nucleic acid sequence encoding" refers to a nucleic acid which directs the expression of a specific protein, peptide or amino acid sequence. The nucleic acid sequences include both the DNA strand sequence that is transcribed into RNA and the RNA sequence that is translated into protein, peptide or amino acid sequence. The nucleic acid sequences include both the full length nucleic acid sequences disclosed herein as well as non-full length sequences derived from the full length protein. It being further understood that the sequence includes the degenerate codons of the native sequence or sequences which may be introduced to provide codon preference in a specific host cell. Consequently, the principles of probe selection and array design can readily be extended to analyze more complex polymorphisms (see EP 730,663). For example, to characterize a triallelic SNP polymorphism, three groups of probes can be designed tiled on the three polymorphic forms as described above. As a further example, to analyze a diallelic polymorphism involving a deletion of a nucleotide, one can tile a first group of probes based

on the undeleted polymorphic form as the reference sequence and a second group of probes based on the deleted form as the reference sequence.

For assays of genomic DNA, virtually any biological convenient tissue sample can be used. Suitable samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. Genomic DNA is typically amplified before analysis. Amplification is usually effected by PCR using primers flanking a suitable fragment e.g., of 50-500 nucleotides containing the locus of the polymorphism to be analyzed. Target is usually labeled in the course of amplification. The amplification product can be RNA or DNA, single stranded or double stranded. If double stranded, the amplification product is typically denatured before application to an array. If genomic DNA is analyzed without amplification, it may be desirable to remove RNA from the sample before applying it to the array. Such can be accomplished by digestion with DNase-free RNase.

DETECTION OF POLYMORPHISMS IN A NUCLEIC ACID SAMPLE

The SNPs disclosed herein can be used to determine which forms of a characterized polymorphism are present in individuals under analysis.

The design and use of allele-specific probes for analyzing polymorphisms is described by e.g., Saiki et al., Nature 324, 163-166 (1986); Dattagupta, EP 235,726, Saiki, WO 89/11548. Allele-specific probes can be designed that hybridize to a segment of target DNA from one individual but do not hybridize to the corresponding segment from another individual due to the presence of different polymorphic forms in the respective segments from the two individuals. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Some probes are designed to hybridize to a segment of target DNA such that the polymorphic site aligns with a central position (e.g., in a 15-mer at the 7 position; in a 16-mer, at either the 7, 8 or 9 position) of the probe. This design of probe achieves good discrimination in hybridization between different allelic forms.

Allele-specific probes are often used in pairs, one member of a pair showing a perfect match to a reference form of a target sequence and the other member showing a perfect match to a variant form. Several pairs of probes can then be immobilized on the same support for simultaneous analysis of multiple polymorphisms within the same target sequence.

The polymorphisms can also be identified by hybridization to nucleic acid arrays, some examples of which are described in published PCT application WO 95/11995. WO 95/11995 also describes subarrays that are optimized for detection of a variant form of a precharacterized polymorphism. Such a subarray contains probes designed to be complementary to a second reference sequence, which is an allelic variant of the first reference sequence. The second group of probes is designed by the same principles, except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group (or further groups) can be particularly useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (e.g., two or more mutations within 9 to 21 bases).

An allele-specific primer hybridizes to a site on a target DNA overlapping a polymorphism and only primes amplification of an allelic form to which the primer exhibits perfect complementarity. See Gibbs, Nucleic Acid Res. 17 2427-2448 (1989). This primer is used in conjunction with a second primer which hybridizes at a distal site. Amplification proceeds from the two-primers, resulting in a detectable product which indicates the particular allelic form is present. A control is usually performed with a second pair of primers, one of which shows a single base mismatch at the polymorphic site and the other of which exhibits perfect complementarity to a distal site. The single-base mismatch prevents amplification and no detectable product is formed. The method works best when the mismatch is included in the 3'-most position of the oligonucleotide aligned with the polymorphism because this position is most destabilizing to elongation from the primer (see, e.g., WO 93/22456).

Amplification products generated using the polymerase chain reaction can be analyzed by the use of denaturing gradient gel electrophoresis. Different alleles can be identified based on the different sequence-dependent melting properties and electrophoretic migration of DNA in solution. Erlich, ed., PCR Technology, Principles and Applications for DNA Amplification, (W.H. Freeman and Co New York, 1992, Chapter 7).

Alleles of target sequences can be differentiated using single-strand conformation polymorphism analysis, which identifies base differences by alteration in electrophoretic migration of single stranded PCR products, as described in Orita et al., Proc. Nat. Acad. Sci. 86, 2766-2770 (1989). Amplified PCR products can be generated and heated or

otherwise denatured, to form single stranded amplification products. Single-stranded nucleic acids may refold or form secondary structures which are partially dependent on the base sequence. The different electrophoretic mobilities of single-stranded amplification products can be related to base-sequence differences between alleles of target sequences.

The genotype of an individual with respect to a pathology suspected of being caused by a genetic polymorphism may be assessed by association analysis. Phenotypic traits suitable for association analysis include diseases that have known but hitherto unmapped genetic components (e.g., agammaglobulinemia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease, familial hypercholesterolemia, polycystic kidney disease, hereditary spherocytosis, von Willebrand's disease, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, osteogenesis imperfecta, and acute intermittent porphyria).

Phenotypic traits also include symptoms of, or susceptibility to, multifactorial diseases of which a component is or may be genetic, such as autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non- independent), systemic lupus erythematosus and Graves disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, oral cavity, ovary, pancreas, prostate, skin, stomach, leukemia, liver, lung, and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness, obesity), strength, speed, endurance, fertility, and susceptibility or receptivity to particular drugs or therapeutic treatments.

Determination of which polymorphic forms occupy a set of polymorphic sites in an individual identifies a set of polymorphic forms that distinguishes the individual. See generally National Research Council, *The Evaluation of Forensic DNA Evidence* (Eds. Pollard et al., National Academy Press, DC, 1996). Since the polymorphic sites are within a 50,000 bp region in the human genome, the probability of recombination between these polymorphic sites is low. That low probability means the haplotype (the set of all 10 polymorphic sites) set forth in this application should be inherited without change for at least several generations. The more sites that are analyzed the lower the probability that the set of polymorphic forms in one individual is the same as that in an unrelated individual. Preferably, if multiple sites are analyzed, the sites are unlinked. Thus, polymorphisms of the

invention are often used in conjunction with polymorphisms in distal genes. Preferred polymorphisms for use in forensics are diallelic because the population frequencies of two polymorphic forms can usually be determined with greater accuracy than those of multiple polymorphic forms at multi-allelic loci.

The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance.

p(ID) is the probability that two random individuals have the same polymorphic or allelic form at a given polymorphic site. In diallelic loci, four genotypes are possible: AA, AB, BA, and BB. If alleles A and B occur in a haploid genome of the organism with frequencies x and y, the probability of each genotype in a diploid organism are (see WO 95/12607):

$$\text{Homozygote: } p(AA)=x^2$$

$$\text{Homozygote: } p(BB)=y^2=(1-x)^2$$

$$\text{Single Heterozygote: } p(AB)=p(BA)=xy=x(1-x)$$

$$\text{Both Heterozygotes: } p(AB+BA)=2xy=2x(1-x)$$

The probability of identity at one locus (i.e, the probability that two individuals, picked at random from a population will have identical polymorphic forms at a given locus) is given by the equation:

$$p(ID)=(x^2)^2+(2xy)^2+(y^2)^2.$$

These calculations can be extended for any number of polymorphic forms at a given locus. For example, the probability of identity p(ID) for a 3-allele system where the alleles have the frequencies in the population of x, y and z, respectively, is equal to the sum of the squares of the genotype frequencies:

$$p(ID)=x^{4+}(2xy)^{2+}(2yz)^{2+}(2xz)^{2+}z^{4+}y^{4}$$

In a locus of n alleles, the appropriate binomial expansion is used to calculate p(ID) and p(exc).

The cumulative probability of identity (cum p(ID)) for each of multiple unlinked loci is determined by multiplying the probabilities provided by each locus:

$$cum\ p(ID)=p(ID1)p(ID2)p(ID3)\ldots p(IDn)$$

The cumulative probability of non-identity for n loci (i.e. the probability that two random individuals will be different at 1 or more loci) is given by the equation:

$$cum\ p(nonID)=1-cum\ p(ID).$$

If several polymorphic loci are tested, the cumulative probability of non-identity for random individuals becomes very high (e.g., one billion to one). Such probabilities can be taken into account together with other evidence in determining the guilt or innocence of the suspect.

The object of paternity testing is usually to determine whether a male is the father of a child. In most cases, the mother of the child is known and thus, the mother's contribution to the child's genotype can be traced. Paternity testing investigates whether the part of the child's genotype not attributable to the mother is consistent with that of the putative father. Paternity testing can be performed by analyzing sets of polymorphisms in the putative father and the child.

If the set of polymorphisms in the child attributable to the father does not match the putative father, it can be concluded, barring experimental error, that the putative father is not the real father. If the set of polymorphisms in the child attributable to the father does match the set of polymorphisms of the putative father, a statistical calculation can be performed to determine the probability of coincidental match.

The probability of parentage exclusion (representing the probability that a random male will have a polymorphic form at a given polymorphic site that makes him incompatible as the father) is given by the equation (see WO 95/12607):

$p(exc)=xy(1-xy)$

where x and y are the population frequencies of alleles A and B of a diallelic polymorphic site. (At a triallelic site p(exc)=xy(1-xy)+ yz(1-yz)+ xz(1-xz)+ 3xyz(1-xyz))), where x, y and z and the respective population frequencies of alleles A, B and C). The probability of non-exclusion is:

$p(non\text{-}exc)=1-p(exc)$

The cumulative probability of non-exclusion (representing the value obtained when n loci are used) is thus:

$cum\ p(non\text{-}exc)=p(non\text{-}exc1)p(non\text{-}exc2)p(non\text{-}exc3)\ldots p(non\text{-}excn)$

The cumulative probability of exclusion for n loci (representing the probability that a random male will be excluded) is:

$cum\ p(exc)=1-cum\ p(non\text{-}exc)$.

If several polymorphic loci are included in the analysis, the cumulative probability of exclusion of a random male is very high. This probability can be taken into account in assessing the liability of a putative father whose polymorphic marker set matches the child's polymorphic marker set attributable to his/her father.

The polymorphisms of the invention may contribute to the phenotype of an organism in different ways. Some polymorphisms occur within a protein coding sequence and contribute to phenotype by affecting protein structure. The effect may be neutral, beneficial or detrimental, or both beneficial and detrimental, depending on the circumstances. For example, a heterozygous sickle cell mutation confers resistance to malaria, but a homozygous sickle cell mutation is usually lethal. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on replication, transcription, and translation. A single polymorphism may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by polymorphisms in different genes. Further, some

polymorphisms predispose an individual to a distinct mutation that is causally related to a certain phenotype.

Phenotypic traits include diseases that have known but hitherto unmapped genetic components. Phenotypic traits also include symptoms of, or susceptibility to, multifactorial diseases of which a component is or may be genetic, such as autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non-independent), systemic lupus erythematosus and Graves disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, leukemia, liver, lung, oral cavity, ovary, pancreas, prostate, skin, stomach and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness, obesity), strength, speed, endurance, fertility, and susceptibility or receptivity to particular drugs or therapeutic treatments.

Correlation is performed for a population of individuals who have been tested for the presence or absence of a phenotypic trait of interest and for polymorphic marker sets. To perform such analysis, the presence or absence of a set of polymorphisms (i.e. a polymorphic set) is determined for a set of the individuals, some of whom exhibit a particular trait, and some of whom exhibit lack of the trait. The alleles of each polymorphism of the set are then reviewed to determine whether the presence or absence of a particular allele is associated with the trait of interest. Correlation can be performed by standard statistical methods and statistically significant correlations between polymorphic form(s) and phenotypic characteristics are noted. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased milk production of a farm animal.

Such correlations can be exploited in several ways. In the case of a strong correlation between a set of one or more polymorphic forms and a disease for which treatment is available, detection of the polymorphic form set in a human or animal patient may justify immediate administration of treatment, or at least the institution of regular monitoring of the patient. Detection of a polymorphic form correlated with serious disease in a couple contemplating a family may also be valuable to the couple in their reproductive decisions. For example, the female partner might elect to undergo in vitro fertilization to avoid the

possibility of transmitting such a polymorphism from her husband to her offspring. In the case of a weaker, but still statistically significant correlation between a polymorphic set and human disease, immediate therapeutic intervention or monitoring may not be justified. Nevertheless, the patient can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little cost to the patient but confer potential benefits in reducing the risk of conditions to which the patient may have increased susceptibility by virtue of variant alleles. Identification of a polymorphic set in a patient correlated with enhanced receptiveness to one of several treatment regimes for a disease indicates that this treatment regime should be followed.

For animals and plants, correlations between characteristics and phenotype are useful for breeding for desired characteristics. For example, Beitz et al., U.S. Pat. No. 5,292,639 discuss use of bovine mitochondrial polymorphisms in a breeding program to improve milk production in cows. To evaluate the effect of mtDNA D-loop sequence polymorphism on milk production, each cow was assigned a value of 1 if variant or 0 if wild type with respect to a prototypical mitochondrial DNA sequence at each of 17 locations considered.

The previous section concerns identifying correlations between phenotypic traits and polymorphisms that directly or indirectly contribute to those traits. The present section describes identification of a physical linkage between a genetic locus associated with a trait of interest and polymorphic markers that are not associated with the trait, but are in physical proximity with the genetic locus responsible for the trait and co-segregate with it. Such analysis is useful for mapping a genetic locus associated with a phenotypic trait to a chromosomal position, and thereby cloning gene(s) responsible for the trait. See Lander et al., *Proc. Natl. Acad. Sci.* (USA) 83, 7353-7357 (1986); Lander et al., *Proc. Natl. Acad. Sci.* (USA) 84, 2363-2367 (1987); Donis-Keller et al., *Cell* 51, 319-337 (1987); Lander et al., *Genetics* 121, 185-199 (1989)). Genes localized by linkage can be cloned by a process known as directional cloning. See Wainwright, *Med. J. Australia* 159, 170-174 (1993); Collins, *Nature Genetics* 1, 3-6 (1992) (each of which is incorporated by reference in its entirety for all purposes).

Linkage studies are typically performed on members of a family. Available members of the family are characterized for the presence or absence of a phenotypic trait and for a set of polymorphic markers. The distribution of polymorphic markers in an informative meiosis is then analyzed to determine which polymorphic markers co-segregate with a phenotypic

trait. See, e.g., Kerem et al., *Science* 245, 1073-1080 (1989); Monaco et al., *Nature* 316, 842 (1985); Yamoka et al., *Neurology* 40, 222-226 (1990); Rossiter et al., *FASEB Journal* 5, 21-27 (1991).

Linkage is analyzed by calculation of LOD (log of the odds) values. A lod value is the relative likelihood of obtaining observed segregation data for a marker and a genetic locus when the two are located at a recombination fraction *RF*, versus the situation in which the two are not linked, and thus segregating independently (Thompson & Thompson, *Genetics in Medicine* (5th ed, W.B. Saunders Company, Philadelphia, 1991); Strachan, "Mapping the human genome" in *The Human Genome* (BIOS Scientific Publishers Ltd, Oxford), Chapter 4). A series of likelihood ratios are calculated at various recombination fractions (*RF*), ranging from *RF*=0.0 (coincident loci) to *RF*=0.50 (unlinked). Thus, the likelihood at a given value of *RF* is: probability of data if loci linked at *RF* to probability of data if loci unlinked. The computed likelihood is usually expressed as the $\log_{10}$ of this ratio (i.e., a lod score). For example, a lod score of 3 indicates 1000:1 odds against an apparent observed linkage being a coincidence. The use of logarithms allows data collected from different families to be combined by simple addition. Computer programs are available for the calculation of lod scores for differing values of *RF* (e.g., LIPED, MLINK (Lathrop, *Proc. Nat. Acad. Sci.* (USA) 81, 3443-3446 (1984)). For any particular lod score, a recombination fraction may be determined from mathematical tables. See Smith et al., *Mathematical tables for research workers in human genetics* (Churchill, London, 1961); Smith, *Ann. Hum. Genet.* 32, 127-150 (1968). The value of *RF* at which the lod score is the highest is considered to be the best estimate of the recombination fraction.

Positive lod score values suggest that the two loci are linked, whereas negative values suggest that linkage is less likely (at that value of *RF*) than the possibility that the two loci are unlinked. By convention, a combined lod score of + 3 or greater (equivalent to greater than 1000:1 odds in favor of linkage) is considered definitive evidence that two loci are linked. Similarly, by convention, a negative lod score of -2 or less is taken as definitive evidence against linkage of the two loci being compared. Negative linkage data are useful in excluding a chromosome or a segment thereof from consideration. The search focuses on the remaining non-excluded chromosomal locations.

The invention further provides transgenic nonhuman animals capable of expressing an exogenous variant gene and/or having one or both alleles of an endogenous variant gene

inactivated. Expression of an exogenous variant gene is usually achieved by operably linking the gene to a promoter and optionally an enhancer, and microinjecting the construct into a zygote. See Hogan et al., "Manipulating the Mouse Embryo, A Laboratory Manual," Cold Spring Harbor Laboratory. (1989). Inactivation of endogenous variant genes can be achieved by forming a transgene in which a cloned variant gene is inactivated by insertion of a positive selection marker. See Capecchi, Science 244, 1288-1292 The transgene is then introduced into an embryonic stem cell, where it undergoes homologous recombination with an endogenous variant gene. Mice and other rodents are preferred animals. Such animals provide useful drug screening systems.

The invention further provides methods for assessing the pharmacogenomic susceptibility of a subject harboring a single nucleotide polymorphism to a particular pharmaceutical compound, or to a class of such compounds. Genetic polymorphism in drug-metabolizing enzymes, drug transporters, receptors for pharmaceutical agents, and other drug targets have been correlated with individual differences based on distinction in the efficacy and toxicity of the pharmaceutical agent administered to a subject. Pharmocogenomic characterization of a subjects susceptibility to a drug enhances the ability to tailor a dosing regimen to the particular genetic constitution of the subject, thereby enhancing and optimizing the therapeutic effectiveness of the therapy.

In cases in which a cSNP leads to a polymorphic protein that is ascribed to be the cause of a pathological condition, method of treating such a condition includes administering to a subject experiencing the pathology the wild type cognate of the polymorphic protein. Once administered in an effective dosing regimen, the wild type cognate provides complementation or remediation of the defect due to the polymorphic protein. The subject's condition is ameliorated by this protein therapy.

A subject suspected of suffering from a pathology ascribable to a polymorphic protein that arises from a cSNP is to be diagnosed using any of a variety of diagnostic methods capable of identifying the presence of the cSNP in the nucleic acid, or of the cognate polymorphic protein, in a suitable clinical sample taken from the subject. Once the presence of the cSNP has been ascertained, and the pathology is correctable by administering a normal or wild-type gene, the subject is treated with a pharmaceutical composition that includes a nucleic acid that harbors the correcting wild-type gene, or a fragment containing a correcting sequence of the wild-type gene. Non-limiting examples of ways in which such a nucleic acid

may be administered include incorporating the wild-type gene in a viral vector, such as an adenovirus or adeno associated virus, and administration of a naked DNA in a pharmaceutical composition that promotes intracellular uptake of the administered nucleic acid. Once the nucleic acid that includes the gene coding for the wild-type allele of the polymorphism is incorporated within a cell of the subject, it will initiate *de novo* biosynthesis of the wild-type gene product. If the nucleic acid is further incorporated into the genome of the subject, the treatment will have long-term effects, providing *de novo* synthesis of the wild-type protein for a prolonged duration. The synthesis of the wild-type protein in the cells of the subject will contribute to a therapeutic enhancement of the clinical condition of the subject.

A subject suffering from a pathology ascribed to a SNP may be treated so as to correct the genetic defect. (See Kren et al., Proc. Natl. Acad. Sci. USA 96:10349-10354 (1999)). Such a subject is identified by any method that can detect the polymorphism in a sample drawn from the subject. Such a genetic defect may be permanently corrected by administering to such a subject a nucleic acid fragment incorporating a repair sequence that supplies the wild-type nucleotide at the position of the SNP. This site-specific repair sequence encompasses an RNA/DNA oligonucleotide which operates to promote endogenous repair of a subject's genomic DNA. Upon administration in an appropriate vehicle, such as a complex with polyethylenimine or encapsulated in anionic liposomes, a genetic defect leading to an inborn pathology may be overcome, as the chimeric oligonucleotides induces incorporation of the wild-type sequence into the subject's genome. Upon incorporation, the wild-type gene product is expressed, and the replacement is propagated, thereby engendering a permanent repair.

The invention further provides kits comprising at least one allele-specific oligonucleotide as described above. Often, the kits contain one or more pairs of allele-specific oligonucleotides hybridizing to different forms of a polymorphism. In some kits, the allele-specific oligonucleotides are provided immobilized to a substrate. For example, the same substrate can comprise allele-specific oligonucleotide probes for detecting at least 10, 100, 1000 or all of the polymorphisms shown in the Table. Optional additional components of the kit include, for example, restriction enzymes, reverse-transcriptase or polymerase, the substrate nucleoside triphosphates, means used to label (for example, an avidin-enzyme conjugate and enzyme substrate and chromogen if the label is biotin), and the appropriate buffers for reverse transcription, PCR, or hybridization reactions. Usually, the kit also contains instructions for carrying out the hybridizing methods.

Several aspects of the present invention rely on having available the polymorphic proteins encoded by the nucleic acids comprising a SNP of the inventions. There are various methods of isolating these nucleic acid sequences. For example, DNA is isolated from a genomic or cDNA library using labeled oligonucleotide probes having sequences complementary to the sequences disclosed herein.

Such probes can be used directly in hybridization assays. Alternatively probes can be designed for use in amplification techniques such as PCR.

To prepare a cDNA library, mRNA is isolated from tissue such as heart or pancreas, preferably a tissue wherein expression of the gene or gene family is likely to occur. cDNA is prepared from the mRNA and ligated into a recombinant vector. The vector is transfected into a recombinant host for propagation, screening and cloning. Methods for making and screening cDNA libraries are well known, See Gubler, U. and Hoffman, B.J. Gene 25:263-269 (1983) and Sambrook et al.

For a genomic library, for example, the DNA is extracted from tissue and either mechanically sheared or enzymatically digested to yield fragments of about 12-20 kb. The fragments are then separated by gradient centrifugation from undesired sizes and are constructed in bacteriophage lambda vectors. These vectors and phage are packaged *in vitro,* as described in Sambrook, et al. Recombinant phage are analyzed by plaque hybridization as described in Benton and Davis, Science 196:180-l 82 (1977). Colony hybridization is carried out as generally described in M. Grunstein et al. Proc. Natl. Acad. Sci. USA. 72:3961-3965 (1975). DNA of interest is identified in either cDNA or genomic libraries by its ability to hybridize with nucleic acid probes, for example on Southern blots, and these DNA regions are isolated by standard methods familiar to those of skill in the art. See Sambrook, et al.

In PCR techniques, oligonucleotide primers complementary to the two 3' borders of the DNA region to be amplified are synthesized. The polymerase chain reaction is then carried out using the two primers. See PCR Protocols: a Guide to Methods and Applications (Innis, M, Gelfand, D., Sninsky, J. and White, T., eds.), Academic Press, San Diego (1990). Primers can be selected to amplify the entire regions encoding a full-length sequence of interest or to amplify smaller DNA segments as desired. PCR can be used in a variety of protocols to isolate cDNAs encoding a sequence of interest. In these protocols, appropriate primers and probes for amplifying DNA encoding a sequence of interest are generated from

analysis of the DNA sequences listed herein. Once such regions are PCR-amplified, they can be sequenced and oligonucleotide probes can be prepared from the sequence.

Once DNA encoding a sequence comprising a cSNP is isolated and cloned, one can express the encoded polymorphic proteins in a variety of recombinantly engineered cells. It is expected that those of skill in the art are knowledgeable in the numerous expression systems available for expression of DNA encoding a sequence of interest. No attempt to describe in detail the various methods known for the expression of proteins in prokaryotes or eukaryotes is made here.

In brief summary, the expression of natural or synthetic nucleic acids encoding a sequence of interest will typically be achieved by operably linking the DNA or cDNA to a promoter (which is either constitutive or inducible), followed by incorporation into an expression vector. The vectors can be suitable for replication and integration in either prokaryotes or eukaryotes. Typical expression vectors contain initiation sequences, transcription and translation terminators, and promoters useful for regulation of the expression of a polynucleotide sequence of interest. To obtain high level expression of a cloned gene, it is desirable to construct expression plasmids which contain, at the minimum, a strong promoter to direct transcription, a ribosome binding site for translational initiation, and a transcription/translation terminator. The expression vectors may also comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the plasmid in both eukaryotes and prokaryotes, i.e., shuttle vectors, and selection markers for both prokaryotic and eukaryotic systems. See Sambrook et al.

A variety of prokaryotic expression systems may be used to express the polymorphic proteins of the invention. Examples include *E. coli*, Bacillus, Streptomyces, and the like.

It is preferred to construct expression plasmids which contain, at the minimum, a strong promoter to direct transcription, a ribosome binding site for translational initiation, and a transcription/translation terminator. Examples of regulatory regions suitable for this purpose in *E. coli* are the promoter and operator region of the *E. coli* tryptophan biosynthetic pathway as described by Yanofsky, C., J. Bacterial. 158:1018-1024 (1984) and the leftward promoter of phage lambda as described by Λ, I. and Hagen, D., Ann. Rev. Genet. 14:399-445 (1980). The inclusion of selection markers in DNA vectors transformed in *E. coli* is also useful. Examples of such markers include genes specifying resistance to ampicillin,

tetracycline, or chloramphenicol. See Sambrook et al. for details concerning selection markers for use in *E. coli.*

To enhance proper folding of the expressed recombinant protein, during purification from *E. coli*, the expressed protein may first be denatured and then renatured. This can be accomplished by solubilizing the bacterially produced proteins in a chaotropic agent such as guanidine HCI and reducing all the cysteine residues with a reducing agent such as beta-mercaptoethanol. The protein is then renatured, either by slow dialysis or by gel filtration. See U.S. Patent No. 4,511,503. Detection of the expressed antigen is achieved by methods known in the art as radioimmunoassay, or Western blotting techniques or immunoprecipitation. Purification from *E. coli* can be achieved following procedures such as those described in U.S. Patent No. 4,511,503.

Any of a variety of eukaryotic expression systems such as yeast, insect cell lines, bird, fish, and mammalian cells, may also be used to express a polymorphic protein of the invention. As explained briefly below, a nucleotide sequence harboring a cSNP may be expressed in these eukaryotic systems. Synthesis of heterologous proteins in yeast is well known. Methods in Yeast Genetics, Sherman, F., et al., Cold Spring Harbor Laboratory, (1982) is a well recognized work describing the various methods available to produce the protein in yeast. Suitable vectors usually have expression control sequences, such as promoters, including 3-phosphogtycerate kinase or other glycolytic enzymes, and an origin of replication, termination sequences and the like as desired. For instance, suitable vectors are described in the literature (Botstein, et al., Gene 8:17-24 (1979); Broach, et al., Gene 8:121-133 (1979)).

Two procedures are used in transforming yeast cells. In one case, yeast cells are first converted into protoplasts using zymolyase, lyticase or glusulase, followed by addition of DNA and polyethylene glycol (PEG). The PEG-treated protoplasts are then regenerated in a 3% agar medium under selective conditions. Details of this procedure are given in the papers by J.D. Beggs, Nature (London) 275:104-109 (1978); and Hinnen, A., et al., Proc. Natl. Acad. Sci. USA, 75:1929-1933 (1978). The second procedure does not involve removal of the cell wall. Instead the cells are treated with lithium chloride or acetate and PEG and put on selective plates (Ito, H., et al., J. Bact, 153163-168 (1983)) cells and applying standard protein isolation techniques to the lysates:.

The purification process can be monitored by using Western blot techniques or radioimmunoassay or other standard techniques. The sequences encoding the proteins of the invention can also be ligated to various immunoassay expression vectors for use in transforming cell cultures of, for instance, mammalian, insect, bird or fish origin. Illustrative of cell cultures useful for the production of the polypeptides are mammalian cells. Mammalian cell systems often will be in the form of monolayers of cells although mammalian cell suspensions may also be used. A number of suitable host cell lines capable of expressing intact proteins have been developed in the art, and include the HEK293, BHK21, and CHO cell lines, and various human cells such as COS cell lines, HeLa cells, myeloma cell lines, Jurkat cells, etc. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter (e.g., the CMV promoter, a HSV *tk* promoter or *pgk* (phosphoglycerate kinase) promoter), an enhancer (Queen et al. Immunol. Rev, 89:49 (1986)) and necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites (e.g., an SV40 large T Ag poly A addition site), and transcriptional terminator sequences.

Other animal cells are available, for instance, from the American Type Culture Collection Catalogue of Cell Lines and Hybridomas (7th edition, (1992)). Appropriate vectors for expressing the proteins of the invention in insect cells are usually derived from baculovirus. Insect cell lines include mosquito larvae, silkworm, armyworm, moth and Drosophila cell lines such as a Schneider cell line (See Schneider J. Embryol. Exp. Morphol., 27:353-365 (1987). As indicated above, the vector, e.g., a plasmid, which is used to transform the host cell, preferably contains DNA sequences to initiate transcription and sequences to control the translation of the protein. These sequences are referred to as expression control sequences. As with yeast, when higher animal host cells are employed, polyadenylation or transcription terminator sequences from known mammalian genes need to be incorporated into the vector. An example of a terminator sequence is the polyadenylation sequence from the bovine growth hormone gene. Sequences for accurate splicing of the transcript may also be included. An example of a splicing sequence is the VP1 intron from SV4O (Sprague, J. et a/., J. Virol. 45: 773-781 (1983)). Additionally, gene sequences to control replication in the host cell may be Saveria-Campo, M., 1985, "Bovine Papilloma virus DNA a Eukaryotic Cloning Vector" in DNA Cloning Vol. II a Practical Approach Ed. D.M. Glover, IRL Press, Arlington, Virginia pp. 213-238. The host cells are competent or rendered competent for transformation by various means. There are several well-known

methods of introducing DNA into animal cells. These include: calcium phosphate precipitation, fusion of the recipient cells with bacterial protoplasts containing the DNA, treatment of the recipient cells with liposomes containing the DNA, DEAE dextran, electroporation and micro-injection of the DNA directly into the cells.

The transformed cells are cultured by means well known in the art (Biochemical Methods in Cell Culture and Virology, Kuchler, R.J., Dowden, Hutchinson and Ross, Inc., (1977)). The expressed polypeptides are isolated from cells grown as suspensions or as monolayers. The latter are recovered by well known mechanical, chemical or enzymatic means.

General methods of expressing recombinant proteins are also known and are exemplified in R. Kaufman, Methods in Enzymology 185, 537-566 (1990). As defined herein "operably linked" refers to linkage of a promoter upstream from a DNA sequence such that the promoter mediates transcription of the DNA sequence. Specifically, "operably linked" means that the isolated polynucleotide of the invention and an expression control sequence are situated within a vector or cell in such a way that the gene encoding the protein is expressed by a host cell which has been transformed (transfected) with the ligated polynucleotide/expression sequence. The term "vector", refers to viral expression systems, autonomous self-replicating circular DNA (plasmids), and includes both expression and nonexpression plasmids.

The term "gene" as used herein is intended to refer to a nucleic acid sequence which encodes a polypeptide. This definition includes various sequence polymorphisms, mutations, and/or sequence variants wherein such alterations do not affect the function of the gene product. The term "gene" is intended to include not only coding sequences but also regulatory regions such as promoters, enhancers, termination regions and similar untranslated nucleotide sequences. The term further includes all introns and other DNA sequences spliced from the mRNA transcript, along with variants resulting from alternative splice sites.

A number of types of cells may act as suitable host cells for expression of the protein. Mammalian host cells include, for example, monkey COS cells, Chinese Hamster Ovary (CHO) cells, human kidney 293 cells, human epidermal A43 1 cells, human Co10205 cells, 3T3 cells, CV-1 cells, other transformed primate cell lines, normal diploid cells, cell strains derived from in vitro culture of primary tissue, primary explants, HeLa cells, mouse L cells,

BHK, HL- 60, U937, HaK or Jurkat cells. Alternatively, it may be possible to produce the protein in lower eukaryotes such as yeast or in prokaryotes such as bacteria. Potentially suitable yeast strains include *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Kluyveromyces* strains, Candida or any yeast strain capable of expressing heterologous proteins. Potentially suitable bacterial strains include *Escherichia coli, Bacillus subtilis, Salmonella typhimurium,* or any bacterial strain capable of expressing heterologous proteins. If the protein is made in yeast or bacteria, it may be necessary to modify the protein produced therein, for example by phosphorylation or glycosylation of the appropriate sites, in order to obtain the functional protein.

The protein may also be produced by operably linking the isolated polynucleotide of the invention to suitable control sequences in one or more insect expression vectors, and employing an insect expression system. Materials and methods for baculovirus/insect cell expression systems are commercially available in kit form from, e.g., Invitrogen, San Diego, California, U.S.A. (the MaxBac© kit), and such methods are well known in the art, as described in Summers and Smith, Texas Agricultural Experiment Station Bulletin No. 1555 (1987), incorporated herein by reference. As used herein, an insect cell capable of expressing a polynucleotide of the present invention is "transformed." The protein of the invention may be prepared by culturing transformed host cells under culture conditions suitable to express the recombinant protein.

The polymorphic protein of the invention may also be expressed as a product of transgenic animals, e.g., as a component of the milk of transgenic cows, goats, pigs, or sheep which are characterized by somatic or germ cells containing a nucleotide sequence encoding the protein. The protein may also be produced by known conventional chemical synthesis. Methods for constructing the proteins of the present invention by synthetic means are known to those skilled in the art.

The polymorphic proteins produced by recombinant DNA technology may be purified by techniques commonly employed to isolate or purify recombinant proteins. Recombinantly produced proteins can be directly expressed or expressed as a fusion protein. The protein is then purified by a combination of cell lysis (e.g., sonication) and affinity chromatography. For fusion products, subsequent digestion of the fusion protein with an appropriate proteolytic enzyme releases the desired polypeptide. The polypeptides of this invention may be purified to substantial purity by standard techniques well known in the art, including

selective precipitation with such substances as ammonium sulfate, column chromatography, immunopurification methods, and others. See, for instance, R. Scopes, Protein Purification: Principles and Practice, Springer-Verlag: New York (1982), incorporated herein by reference. For example, in an embodiment, antibodies may be raised to the proteins of the invention as described herein. Cell membranes are isolated from a cell line expressing the recombinant protein, the protein is extracted from the membranes and immunoprecipitated. The proteins may then be further purified by standard protein chemistry techniques as described above.

The resulting expressed protein may then be purified from such culture (i.e., from culture medium or cell extracts) using known purification processes, such as gel filtration and ion exchange chromatography. The purification of the protein may also include an affinity column containing agents which will bind to the protein; one or more column steps over such affinity resins as concanavalin A-agarose, heparin-Toyopearl@ or Cibacrom blue 3GA Sepharose B; one or more steps involving hydrophobic interaction chromatography using such resins as phenyl ether, butyl ether, or propyl ether; or immunoaffinity chromatography. Alternatively, the protein of the invention may also be expressed in a form which will facilitate purification. For example, it may be expressed as a fusion protein, such as those of maltose binding protein (MBP), glutathione-S-transferase (GST) or thioredoxin (TRX). Kits for expression and purification of such fusion proteins are commercially available from New England BioLab (Beverly, MA), Pharmacia (Piscataway, NJ) and InVitrogen, respectively. The protein can also be tagged with an epitope and subsequently purified by using a specific antibody directed to such epitope. One such epitope ("Flag") is commercially available from Kodak (New Haven, CT). Finally, one or more reverse-phase high performance liquid chromatography (RP- HPLC) steps employing hydrophobic RP-HPLC media, e.g., silica gel having pendant methyl or other aliphatic groups, can be employed to further purify the protein. Some or all of the foregoing purification steps, in various combinations, can also be employed to provide a substantially homogeneous isolated recombinant protein. The protein thus purified is substantially free of other mammalian proteins and is defined in accordance with the present invention as an "isolated protein."

The term "antibody" as used herein refers to immunoglobulin molecules and immunologically active portions of immunoglobulin molecules, *i.e.*, molecules that contain an antigen binding site that specifically binds (immunoreacts with) an antigen, such as polymorphic. Such antibodies include, but are not limited to, polyclonal, monoclonal,

chimeric, single chain, F_{ab} and $F_{(ab')2}$ fragments, and an F_{ab} expression library. In a specific embodiment, antibodies to human polymorphic proteins are disclosed.

The phrase "specifically binds to", "immunospecifically binds to" or is "specifically immunoreactive with", an antibody when referring to a protein or peptide, refers to a binding reaction which is determinative of the presence of the protein in the presence of a heterogeneous population of proteins and other biological materials. Thus, for example, under designated immunoassay conditions, the specified antibodies bind to a particular protein and do not bind in a significant amount to other proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular protein. Of particular interest in the present invention is an antibody that binds immunospecifically to a polymorphic protein but not to its cognate wild type allelic protein, or vice versa. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays are routinely used to select monoclonal antibodies specifically immunoreactive with a protein. See Harlow and Lane (1988) Antibodies, a Laboratory Manual, Cold Spring Harbor Publications, New York, for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity.

Polyclonal and/or monoclonal antibodies that immunospecifically bind to polymorphic gene products but not to the corresponding prototypical or "wild-type" gene products are also provided. Antibodies can be made by injecting mice or other animals with the variant gene product or synthetic peptide. Monoclonal antibodies are screened as are described, for example, in Harlow & Lane, Antibodies, A Laboratory Manual, Cold Spring Harbor Press, New York (1988); Goding, Monoclonal antibodies, Principles and Practice (2d ed.) Academic Press, New York (1986). Monoclonal antibodies are tested for specific immunoreactivity with a variant gene product and lack of immunoreactivity to the corresponding prototypical gene product.

An isolated polymorphic protein, or a portion or fragment thereof, can be used as an immunogen to generate the antibody that binds the polymorphic protein using standard techniques for polyclonal and monoclonal antibody preparation. The full-length polymorphic protein can be used or, alternatively, the invention provides antigenic peptide fragments of polymorphic for use as immunogens. The antigenic peptide of a polymorphic protein of the

invention comprises at least 8 amino acid residues of the amino acid sequence encompassing the polymorphic amino acid and encompasses an epitope of the polymorphic protein such that an antibody raised against the peptide forms a specific immune complex with the polymorphic protein. Preferably, the antigenic peptide comprises at least 10 amino acid residues, more preferably at least 15 amino acid residues, even more preferably at least 20 amino acid residues, and most preferably at least 30 amino acid residues. Preferred epitopes encompassed by the antigenic peptide are regions of polymorphic that are located on the surface of the protein, *e.g.*, hydrophilic regions.

For the production of polyclonal antibodies, various suitable host animals (*e.g.*, rabbit, goat, mouse or other mammal) may be immunized by injection with the polymorphic protein. An appropriate immunogenic preparation can contain, for example, recombinantly expressed polymorphic protein or a chemically synthesized polymorphic polypeptide. The preparation can further include an adjuvant. Various adjuvants used to increase the immunological response include, but are not limited to, Freund's (complete and incomplete), mineral gels (*e.g.*, aluminum hydroxide), surface active substances (*e.g.*, lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, dinitrophenol, etc.), human adjuvants such as *Bacille Calmette-Guerin* and *Corynebacterium parvum*, or similar immunostimulatory agents. If desired, the antibody molecules directed against polymorphic proteins can be isolated from the mammal (*e.g.*, from the blood) and further purified by well known techniques, such as protein A chromatography, to obtain the IgG fraction.

The term "monoclonal antibody" or "monoclonal antibody composition", as used herein, refers to a population of antibody molecules that originates from the clone of a singly hybridoma cell, and that contains only one type of antigen binding site capable of immunoreacting with a particular epitope of a polymorphic protein. A monoclonal antibody composition thus typically displays a single binding affinity for a particular polymorphic protein with which it immunoreacts. For preparation of monoclonal antibodies directed towards a particular polymorphic protein, or derivatives, fragments, analogs or homologs thereof, any technique that provides for the production of antibody molecules by continuous cell line culture may be utilized. Such techniques include, but are not limited to, the hybridoma technique (see Kohler & Milstein, 1975 *Nature* 256: 495-497); the trioma technique; the human B-cell hybridoma technique (see Kozbor, *et al.*, 1983 *Immunol Today* 4: 72) and the EBV hybridoma technique to produce human monoclonal antibodies (see Cole, *et al.*, 1985 In: MONOCLONAL ANTIBODIES AND CANCER THERAPY, Alan R. Liss, Inc., pp.

77-96). Human monoclonal antibodies may be utilized in the practice of the present invention and may be produced by using human hybridomas (see Cote, *et al.*, 1983. *Proc Natl Acad Sci USA* 80: 2026-2030) or by transforming human B-cells with Epstein Barr Virus *in vitro* (see Cole, *et al.*, 1985 In: MONOCLONAL ANTIBODIES AND CANCER THERAPY, Alan R. Liss, Inc., pp. 77-96).

According to the invention, techniques can be adapted for the production of single-chain antibodies specific to a polymorphic protein (see *e.g.*, U.S. Patent No. 4,946,778). In addition, methodologies can be adapted for the construction of F_{ab} expression libraries (see *e.g.*, Huse, *et al.*, 1989 *Science* 246: 1275-1281) to allow rapid and effective identification of monoclonal F_{ab} fragments with the desired specificity for a polymorphic protein or derivatives, fragments, analogs or homologs thereof. Non-human antibodies can be "humanized" by techniques well known in the art. See *e.g.*, U.S. Patent No. 5,225,539. Antibody fragments that contain the idiotypes to a polymorphic protein may be produced by techniques known in the art including, but not limited to: (*i*) an $F_{(ab')2}$ fragment produced by pepsin digestion of an antibody molecule; (*ii*) an F_{ab} fragment generated by reducing the disulfide bridges of an $F_{(ab')2}$ fragment; (*iii*) an F_{ab} fragment generated by the treatment of the antibody molecule with papain and a reducing agent and (*iv*) F_v fragments.

Additionally, recombinant anti-polymorphic protein antibodies, such as chimeric and humanized monoclonal antibodies, comprising both human and non-human portions, which can be made using standard recombinant DNA techniques, are within the scope of the invention. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art, for example using methods described in PCT International Application No. PCT/US86/02269; European Patent Application No. 184,187; European Patent Application No. 171,496; European Patent Application No. 173,494; PCT

In one embodiment, methodologies for the screening of antibodies that possess the desired specificity include, but are not limited to, enzyme-linked immunosorbent assay (ELISA) and other immunologically-mediated techniques known within the art.

Anti-polymorphic protein antibodies may be used in methods known within the art relating to the detection, quantitation and/or cellular or tissue localization of a polymorphic protein (*e.g.*, for use in measuring levels of the polymorphic protein within appropriate physiological samples, for use in diagnostic methods, for use in imaging the protein, and the like). In a given embodiment, antibodies for polymorphic proteins, or derivatives, fragments, analogs or homologs thereof, that contain the antibody-derived CDR, are utilized as pharmacologically-active compounds in therapeutic applications intended to treat a pathology in a subject that arises from the presence of the cSNP allele in the subject.

An anti-polymorphic protein antibody (*e.g.*, monoclonal antibody) can be used to isolate polymorphic proteins by a variety of immunochemical techniques, such as immunoaffinity chromatography or immunoprecipitation. An anti-polymorphic protein antibody can facilitate the purification of natural polymorphic protein from cells and of recombinantly produced polymorphic proteins expressed in host cells. Moreover, an anti-polymorphic protein antibody can be used to detect polymorphic protein (*e.g.*, in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the polymorphic protein. Anti-polymorphic antibodies can be used diagnostically to monitor protein levels in tissue as part of a clinical testing procedure, *e.g.*, to, for example, determine the efficacy of a given treatment regimen. Detection can be facilitated by coupling (*i.e.*, physically linking) the antibody to a detectable substance. Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, -g a lactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H.

EQUIVALENTS

From the foregoing detailed description of the specific embodiments of the invention, it should be apparent that unique compositions and methods of use thereof in SNPs in known genes have been described. Although particular embodiments have been disclosed herein in detail, this has been done by way of example for purposes of illustration only, and is not intended to be limiting with respect to the scope of the appended claims which follow. In particular, it is contemplated by the inventor that various substitutions, alterations, and modifications may be made to the invention without departing from the spirit and scope of the invention as defined by the claims.

479	cg41084924	1422	CCTCCCTCAAGACCATGAGCCGTAG[G/A]AAGCTCTCCCAGCAGAAGGAGAAGA	G	A	Arg	Arg	SILENT-CODING	tm7	Human Gene SWISSPROT-ID:P14416 D(2) DOPAMINE RECEPTOR - HOMO SAPIENS (HUMAN), 443 aa.	1.70E-241	11
480	cg43264978	1175	CCAAGCTCATCGATGCCTCCAGAGT[C/G]TCAGAGACGGAGTACTCTGCCTTGG	C	G	Val	Val	SILENT-CODING	tm7	Human Gene TREMBLNEW-ID:G2736282 G PROTEIN COUPLED RECEPTOR - HOMO SAPIENS (HUMAN), 362 aa.	1.40E-196	
481	cg43264978	1193	CCAGAGTCTCAGAGACGGAGTACTC[T/C]GCCTTGGAGCAGAACACCAAATGAT	T	C	Ser	Ser	SILENT-CODING	tm7	Human Gene TREMBLNEW-ID:G2736282 G PROTEIN COUPLED RECEPTOR - HOMO SAPIENS (HUMAN), 362 aa.	1.40E-196	
482	cg43264978	140	CCCGCCTCAGAACGATGGATCTGCA[C/T]CTCTTCGACTACTCAGAGCCAGGGA	C	T	His	His	SILENT-CODING	tm7	Human Gene TREMBLNEW-ID:G2736282 G PROTEIN COUPLED RECEPTOR - HOMO SAPIENS (HUMAN), 362 aa.	1.40E-196	
483	cg43264978	164	ACCTCTTCGACTACTCAGAGCCAGG[G/C]AACTTCTCGGACATCAGCTGGCCAT	G	C	Gly	Gly	SILENT-CODING	tm7	Human Gene TREMBLNEW-ID:G2736282 G PROTEIN COUPLED RECEPTOR - HOMO SAPIENS (HUMAN), 362 aa.	1.40E-196	
484	cg3001696	1274	GCCCAGACCCCAGCAGCTTCAGCCG[G/C]CCCCGCGAAGCCACGGCCCGCGAGC	G	C	Arg	Arg	SILENT-CODING	tm7	Human Gene SWISSPROT-ID:P41143 DELTA-TYPE OPIOID RECEPTOR (DOR-1) - HOMO SAPIENS (HUMAN), 372 aa.	2.10E-195	1 (1p36.1)
485	cg43967090	790	GGGCTAAATATTTTATGGTTTTATT[C/T]ATTTACTGTGTTCTCATGCTGTGTT	C	T	Phe	Phe	SILENT-CODING	tm7	Human Gene SWISSNEW-ID:O43194 PUTATIVE G PROTEIN-COUPLED RECEPTOR GPR39 - HOMO SAPIENS (HUMAN), 453 aa.	5.10E-195	2

7191	cg43300806	1000	CCAAGGCCAGC CGCAGCTCTGAG AA[G/gap]TCGCT GGCGCTGCTCAA GACCGTAA	G	gap	Ser	Arg (9362)	FRAMES HIFT	tm7	Human Gene SWISSPROT-ID:P21453 PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1 - HOMO SAPIENS (HUMAN), 381 aa.|pcls:TREMBLNEW-ID:G2668608 G PROTEIN-COUPLED RECEPTOR - HOMO SAPIENS (HUMAN), 381 aa (fragment).	9.4E-200	22 (22q13)
7192	cg43967090	1249	TTTCTTAAGCAC TTTTCAGAGCGA G[G/gap]CCGAGC CCCAGTCTAAGT CCCAGTC	G	gap	Ala	Pro (9363)	FRAMES HIFT	tm7	Human Gene SWISSNEW-ID:O43194 PUTATIVE G PROTEIN-COUPLED RECEPTOR GPR39 - HOMO SAPIENS (HUMAN), 453 aa.	5.1E-195	2
7193	cg43967090	923	CAGATTCGGAGG ATCATGGCTGCG G[C/gap]CAAACC CAAGCACGACTG GACGAGG	C	gap	Ala	Ala (9364)	FRAMES HIFT	tm7	Human Gene SWISSNEW-ID:O43194 PUTATIVE G PROTEIN-COUPLED RECEPTOR GPR39 - HOMO SAPIENS (HUMAN), 453 aa.	5.1E-195	2
7194	cg43967090	924	AGATTCGGAGGA TCATGGCTGCGG C[C/gap]AAACCC AAGCACGACTGG ACGAGGT	C	gap	Lys	Asn (9365)	FRAMES HIFT	tm7	Human Gene SWISSNEW-ID:O43194 PUTATIVE G PROTEIN-COUPLED RECEPTOR GPR39 - HOMO SAPIENS (HUMAN), 453 aa.	5.1E-195	2
7195	cg42908704	924	CGCGGGAGGAG GTCAGCAGGACA AG[gap/A]GTGCG GGGGCCGCAAG GATAGCAAG	gap	A	Arg	Arg (9366)	FRAMES HIFT	tm7	Human Gene SWISSPROT-ID:P46663 B1 BRADYKININ RECEPTOR (BK-1 RECEPTOR) - HOMO SAPIENS (HUMAN), 353 aa.	7E-188	14 (14q32.1)
7196	cg43040271	1239	TACGTGAACAAG AGGACGCCCCG GC[G/gap]CGCCG CTGCGCTCATCT CGCTCACT	G	gap	Arg	Pro (9367)	FRAMES HIFT	tm7	Human Gene Similar to SWISSPROT-ID:Q25322 TYRAMINE/OCTOPAMINE RECEPTOR 2 (TYR-LOC 2) - LOCUSTA MIGRATORIA (MIGRATORY LOCUST), 484 aa.|pcls:SPTREMBL-ID:Q25322 GCR2 (G PROTEIN-COUPLED RECEPTOR) - LOCUSTA MIGRATORIA (MIGRATORY LOCUST), 484 aa.	2.9E-74	

<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43967090

<400> 485
gggctaaata ttttatggtt ttatttattt actgtgttct catgctgtgt t 51

<210> 486
<211> 51
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43326635

<400> 486
acttcaactt ctttgtgtgg gtgctccccc cgcttctcct catggtcctc a 51

<210> 487
<211> 51
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg40245117

<400> 487
catgccaatt tgtttccgtc atgaggatgg actacatggt atacttcagc t 51

<210> 488
<211> 51
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg21411454

<400> 488

<211> 51
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg36988276

<400> 7190
ttcaatggaa cccaactaga tgcagctgaa tctaagcgat aataataatt t 51

<210> 7191
<211> 50
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (25)...(26)
<223> Nucleotide deleted between bases 25 and 26

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43300806

<400> 7191
ccaaggccag ccgcagctct gagaatcgct ggcgctgctc aagaccgtaa 50

<210> 7192
<211> 50
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (25)...(26)
<223> Nucleotide deleted between bases 25 and 26

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43967090

<400> 7192
tttcttaagc acttttcaga gcgagccgag ccccagtcta agtcccagtc 50

<210> 7193
<211> 50
<212> DNA
<213> Homo sapiens

<220>

<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (25)...(26)
<223> Nucleotide deleted between bases 25 and 26

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43967090

<400> 7193
cagattcgga ggatcatggc tgcggcaaac ccaagcacga ctggacgagg 50

<210> 7194
<211> 50
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (25)...(26)
<223> Nucleotide deleted between bases 25 and 26

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg43967090

<400> 7194
agattcggag gatcatggct gcggcaaacc caagcacgac tggacgaggt 50

<210> 7195
<211> 51
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature
<222> (0)...(0)
<223> Accession number cg42908704

<400> 7195
cgcgggagga ggtcagcagg acaagagtgc gggggccgca aggatagcaa g 51

<210> 7196
<211> 50
<212> DNA
<213> Homo sapiens

<220>
<221> allele
<222> (26)...(0)
<223> single nucleotide polymorphism

<221> misc_feature